Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the article:

Peña Irles, Víctor; Martín Mor, Adrià. Entrenament de motors de traducció automàtica estadística especialitzats en farmàcia i medicina entre el castellà i el romanés. 2017. 88 p.

This version is available at https://ddd.uab.cat/record/178210

under the terms of the BY-NC-ND license

Entrenament de motors de traducció automàtica estadística entre el castellà i el romanés especialitzats en farmàcia i medicina

Treball de fi de màster

Autor: Víctor Peña Irles
Director: Adrià Martín Mor
Màster en Tradumàtica: Tecnologies de la Traducció
Facultat de Traducció i Interpretació, UAB.

Dades del TFM

Títol (ca): *Entrenament de motors de traducció automàtica estadística especialitzats en farmàcia i medicina entre el castellà i el romanés.*

Título (es): *Entrenamiento de motores de traducción automática estadística especializados en farmacia y medicina entre el castellano y el rumano.*

Title (en): *Training statistical machine translation engines in the pharmaceutical and medical domain between the Romanian and Spanish languages.*

Autor: Víctor Peña Irles

Tutor: Adrià Martín Mor

Centre: Facultat de Traducció i Interpretació

Estudis: Màster en Tradumàtica: Tecnologies de la Traducció

Curs acadèmic: 2016-2017

Paraules clau (ca)

Traducció automàtica, traducció automàtica estadística, castellà, romanès, Moses, MTradumàtica.

Resum (ca)

En aquest TFM es duu a terme l'entrenament de motors de traducció estadística entre el romanés i el castellà, especialitzats en el camp de la farmàcia i la medicina, a la plataforma MTradumàtica. S'hi ofereix una explicació teòrica sobre els sistemes de traducció automàtica existents i, en particular i amb més profunditat, sobre la traducció automàtica estadística i els models que hi entren en joc. Així mateix, s'hi explica amb detall cada un dels passos seguits per a l'entrenament del motor, des de la cerca de recursos lingüístics, la preparació dels materials i la conversió a formats acceptats, fins a l'entrenament mateix del traductor a la plataforma. Una vegada entrenats els motors, s'hi analitza la qualitat de la traducció automàtica dels motors amb avaluadors automàtics i es compara amb altres motors existents. A més a més, s'estudia quin efecte tenen les particularitats del romanés en el resultat del motor, tant pel que fa a la morfologia com a la codificació als dispositius electrònics.

Palabras clave (es)

Traducción automática, traducción automática estadística, rumano, español, Moses, MTradumàtica.

Resumen (es)

En este TFM se lleva a cabo el entrenamiento de motores de traducción estadística entre el rumano y el castellano, especializados en el campo de la farmacia y la medicina, en la plataforma MTradumàtica. Se ofrece una explicación teórica sobre los sistemas de traducción automática existentes y, en particular y con más profundidad, sobre la traducción automática estadística y los modelos que la componen. Asimismo, se explica en detalle cada uno de los pasos seguidos para el entrenamiento del motor, desde la búsqueda de recursos lingüísticos, la preparación de los materiales y la conversión a formatos aceptados, hasta el propio entrenamiento del traductor en la plataforma. Una vez se han entrenado los motores, se analiza la calidad de la traducción automática de los motores con evaluadores automáticos y se compara el resultado con otros motores existentes. Además, se estudia cómo influyen las particularidades del rumano en el resultado del motor, tanto en relación con la morfología como con su codificación en los dispositivos electrónicos.

Keywords (en)

Machine Translation, Statistical Machine Translation, Romanian, Spanish, Moses, MTradumàtica.

Abstract (en)

The aim of this Master's Degree Dissertation is training statistical translation engines in the pharmaceutical and medical domain between the Romanian and Spanish languages on the MTradumàtica platform. A theoretical explanation about the existing machine translation systems is provided and, particularly and more thoroughly, about the statistical machine translation and the models involved in the translation process. Likewise, every single step aiming at the engine training is explained in detail, from the linguistic resources research to the very same translation engine training on the platform, including the material preparation and its conversion into accepted formats. Once the engines have been trained, the quality of the machine translation is analysed through automatic evaluation metrics. The result is compared then with other existing engines. Furthermore, the peculiarities of the Romanian language and the way they affect the engine results have also been studied, both in relation with the morphology and the encoding of the characters on electronic devices.

Índex

Índex d'il·lustracions

1. Introducció

Redactar un treball de fi de màster és, per mi, una ocasió que cal aprofitar per a dedicar-lo a l'estudi d'una temàtica que, a part de tenir un interés comú o general en l'àmbit de les tecnologies de la traducció, puga satisfer les meues inquietuds i interessos personals. La traducció automàtica representa un d'aquests interessos per diverses raons: la revolució tecnològica —associada al fenomen de la globalització, que vivim d'unes quantes dècades ençà i amb més intensitat els últims anys— no sols té un impacte en el desenvolupament de tecnologies de la traducció, com ja s'ha vist en l'arribada i l'ús de memòries de traducció o la traducció automàtica, sinó que alhora aquestes tecnologies són cada vegada més necessàries per a donar resposta als consumidors de traduccions, els quals exigeixen terminis cada vegada més reduïts i tarifes més baixes. La traducció automàtica és i serà, doncs, un component clau que està canviant els paradigmes del sector de la traducció. A més a més, aquestes matèries han estat, segons el meu parer, de les més interessants que he cursat durant el màster de Tradumàtica, tot i que el nombre d'hores dedicades a l'estudi teòric és més aïna baix.

Tanmateix, les exigències dels consumidors de traduccions i els esforços dels proveïdors lingüístics semblen centrar-se tan sols en llengües majoritàries, per a les quals hi ha més demanda. Tenint en compte això, m'agradaria aprofitar els meus coneixements de romanés en aquest treball: es tracta d'una llengua romànica, poc coneguda a casa nostra, malgrat el nombre elevat de romanesos que hi viuen i la proximitat lingüística que manté amb altres llengües llatines com el castellà i el català, en especial. D'altra banda, hi ha una quantitat notable de corpus lingüístics bilingües castellà-romanés d'institucions europees disponibles de manera lliure a internet.

Per tot això, he considerat un bon exercici entrenar un motor de traducció automàtica estadística castellà-romanés. Com que ja existeixen traductors automàtics en aquesta combinació d'idiomes, l'especialitzaré en el camp de la farmàcia i la medicina. Personalment, m'interessa conéixer amb més profunditat el funcionament de la traducció automàtica estadística, com també donar utilitat a recursos lingüístics infrautilitzats, com els corpus bilingües en llengües de menor circulació, com ara el romanés.

Així, aquest treball té un doble interés: d'una banda, descriu amb detall les passes per a l'entrenament d'un motor de traducció automàtica estadística i, de l'altra, aporta un motor de traducció automàtica en un camp d'especialitzat concret en una llengua amb un nombre de parlants baix com el romanés.

2. Objectius: preguntes i hipòtesis

L'objectiu principal del treball és entrenar dos motors de traducció estadística —romanés-castellà i castellà-romanés— especialitzats en medicina i farmàcia, per tal de comprendre amb més detall el funcionament de la TAE i descriure'n la metodologia emprada i l'experiència viscuda en cada pas de l'entrenament. En aquest cas, s'ha utilitzat la plataforma MTradumàtica, del grup de recerca de Tradumàtica del Departament de Traducció i Interpretació i d'Estudis de l'Àsia Oriental de la UAB, basada en la plataforma Moses i desenvolupada per l'empresa Prompsit Language Engineering, especialitzada en tecnologies de la llengua. Al marge dels objectius exposats, a parer meu, l'entrenament del motor és un bon pretext per a estudiar un seguit d'aspectes molt diversos que hi estan relacionats i que apareixen a les diferents etapes de l'entrenament del motor. A continuació, es descriuran aquests objectius concrets per a cada etapa seguida: les fases prèvies, l'entrenament mateix i les fases posteriors.

A les fases prèvies a l'entrenament del motor, un dels objectius principals és la cerca, descàrrega i tractament de recursos lingüístics per a l'entrenament del motor. En aquest cas, cal trobar i seleccionar recursos lingüístics en línia, com ara corpus, del camp de la medicina i la farmàcia, tant bilingües com monolingües, en romanés i en castellà per tal d'entrenar el motor. Així, en aquesta etapa es podrà estimar quina és la quantitat i la disponibilitat de recursos lingüístics en romanés i castellà a la xarxa. A continuació, depenent de la quantitat i la tipologia dels recursos trobats, caldrà trobar com accedir als continguts i utilitzar-los. Així, a banda de determinar la manera més convenient de descarregar grans quantitats de fitxers, podria ser necessari recórrer a tècniques d'extracció de continguts del web —conegudes com a *web scraping*. El tercer objectiu d'aquesta fase serà la preparació dels materials obtinguts i la conversió a un format compatible amb el motor de traducció automàtica triat. En aquesta part, més aïna relacionada amb l'enginyeria lingüística, caldrà examinar en quin format es presenten les dades obtingudes, en quin format les accepta MTradumàtica i trobar una manera de convertir la informació.

Pel que fa a la fase de l'entrenament mateix del motor, hi ha dos aspectes rellevants. D'una banda, cal investigar i descriure, des d'un punt de vista teòric, com funciona un motor de traducció automàtica estadística, tot descrivint els diferents models estadístics que combina perquè funcione. De

l'altra banda, s'explicaran, des d'un punt de vista pràctic, els passos seguits per a l'entrenament mateix del motor a la plataforma MTradumàtica amb els materials obtinguts, part que, pel seu caràcter pràctic, es podria concebre com una mena de manual per a l'entrenament d'aquesta plataforma.

En tercer lloc, a les fases posteriors a l'entrenament del motor, m'interessa analitzar els resultats del motor de traducció automàtica especialitzat en medicina i farmàcia amb objectius diferents. El primer dels objectius consisteix a donar exemples pràctics que complementen l'explicació teòrica dels models estadístics, per tal de descriure com tracta la segmentació del text de partida, les paraules desconegudes, quines serien les traduccions més probables, etc. D'altra banda, m'interessa avaluar la qualitat de la traducció automàtica del motor mitjançant avaluadors automàtics de la qualitat i comparar-la amb la d'altres motors de traducció automàtica disponibles de manera lliure a la xarxa, com ara el traductor de Google, Yandex o Trautorom —el motor de traducció ro>es d'Apertium, basat en regles. Per acabar, també és rellevant analitzar com tracta certes particularitats morfosintàctiques del romanés, com ara les declinacions i articulació per mitjà de sufixos, per a la qual cosa caldrà fer una introducció teòrica breu d'aquestes particularitats.

Pel que fa a les hipòtesis, previsiblement, la cerca d'una gran quantitat de recursos lingüístics bilingües o en romanés —si més no, de fàcil accés i ús— pot ser complicada, més enllà d'aquells recursos pertanyents a institucions europees o internacionals. D'altra banda, reprenent els aspectes lingüístics del romanés, preveig que el motor tindrà dificultats per a resoldre les característiques morfològiques del romanés esmentades. Per acabar, és important destacar que els meus motors seran especialitzats i la resta són genèrics. Malgrat això, probablement serà difícil arribar a obtenir una qualitat comparable a la del traductor de Google, per la quantitat de recursos de què disposen per a l'entrenament del motor, tot i que el fet que el meu motor estiga especialitzat en un camp concret, el de la medicina i la farmàcia pot ajudar a obtenir resultats d'una qualitat mitjanament bona en aquest domini.

3. Marc teòric

En aquest apartat s'estableix el marc teòric del treball. En prime lloc, es farà una introducció sobre els diferents sistemes de traducció automàtica emprats avui dia, com també de les aplicacions i el flux de treball. A continuació, s'aprofundirà en la traducció automàtica estadística i en el funcionament teòric d'aquest sistema. També s'hi dedica un apartat a MTradumàtica, la plataforma emprada per a l'entrenament dels motors i, per acabar, s'expliquen les particularitats del romanés que poden tenir influència en l'entrenament.

3.1. Traducció automàtica

Definició

Una de les definicions més clares de la traducció automàtica la donen Ginestí i Forcada:

> «La traducció automàtica (TA) és el procés de traducció, mitjançant un sistema informàtic (compost per ordinadors i programes), de textos informatitzats escrits en la llengua origen a textos informatitzats escrits en la llengua meta. Un text informatitzat és un fitxer d'ordinador que conté un text en algun format conegut» (Ginestí i Forcada 2009, 43)

Aplicacions de la traducció automàtica

La traducció automàtica té dues aplicacions principals. La més pràctica, anomenada *assimilació* (Ginestí i Forcada 2009), és aquella per mitjà de la qual s'obtenen traduccions per a la comprensió superficial d'un text, sense que n'importe gaire la qualitat o en siga necessària una d'òptima. Un exemple d'assimilació podria ser la traducció automàtica instantània de pàgines web que ofereixen plataformes com Google. D'altra banda, l'aplicació més professional de la traducció automàtica s'anomena *disseminació* (Forcada 2012, 117) i permet l'obtenció de traduccions que requereixen l'actuació humana —postedició— per a aconseguir-ne una qualitat plena. Aquest tipus d'aplicació és un recurs cada vegada més utilitzat per proveïdors de serveis lingüístics, grans empreses i traductors autònoms, atés que veuen incrementada la productivitat: les traduccions es poden obtenir en un termini més curt i amb un preu molt inferior. D'altres autors, com ara Philipp

Koehn, afigen a aquesta classificació una tercera categoria, la *comunicació* (Koehn 2010, 20), que inclou la traducció de correus electrònics, missatgeries o xats, encara que, per les seues característiques, podria pertànyer a la categoria de l'assimilació.

La traducció automàtica és un recurs molt estés en la traducció a gran escala, de manera que grans empreses amb necessitats constants de traducció, tant proveïdores de serveis lingüístics com empreses de sectors concrets, entrenen motors de traducció automàtica especialitzats en un domini o camp lingüístic per tal de millorar la qualitat dels resultats i adaptar-los a les necessitats del client.

Flux de treball

Quant al flux de treball en la disseminació, el primer pas —encara que no sempre es du a terme— és la preedició del text original, és a dir, l'edició prèvia del text de partida basada en l'ús un llenguatge controlat —sense ambigüitats semàntiques, amb una sintaxi simple, etc.— per tal que el motor de traducció automàtica done millors resultats. A la base de coneixement de TAUS donen una bona definició de la preedició:

> «Pre-editing is the process of making the source text amenable to Machine translation by making use of style guides and controlled terminology while applying controlled-language rules in order to improve the output. Pre-editing focuses on modifying input sentences in order to prevent predictable problems usually encountered by MT-systems.» (TAUS 2017).

Una vegada el motor de TA fa el seu treball, és necessària la postedició dels resultats, és a dir, el tractament humà del text traduït per tal d'obtenir-ne una qualitat adequada. Pel que fa a la postedició, n'hi ha dos tipus bàsics: la postedició parcial (o *light post-editing*) i la postedició completa (o *full post-editing).* La primera té l'objectiu de fer les modificacions bàsiques per tal que el text es puga comprendre:

> «This involves taking the raw MT output and performing as few modifications as possible to the text in order to make the translation understandable, factually accurate, and grammatically correct» (Densmer 2014)

L'altre tipus de postedició, la postedició completa, implica l'obtenció d'una qualitat plena del text final, amb la terminologia adequada, coherència, cohesió i sense variació d'estils i sense cap mena d'errada gramatical, com si estigués escrit en la llengua de destí:

> «After this edit, the translation should read as if written in the target language» (Densmer 2014).

D'altra banda, pel que fa als recursos tècnics emprats per a la postedició (PE), se'n poden distingir tres tipus:

- *PE clàssica:* el posteditor treballa amb el text original i la traducció automàtica, confrontats i normalment dividits en segments. És el tipus de postedició més senzill pel que fa als requisits tècnics. Atés que «només cal un programa que presenti d'una manera confrontada aquests segments: el segment original i el corresponent de la traducció en brut» (Martín Mor, Piqué i Huerta, i Sánchez-Gijón 2016, 66).
- *PE integrada*: la postedició es duu a terme en una eina TAO que integre TA i memòries de traducció, per tal d'«aprofitar al màxim els segments de traduccions prèvies que ja han estat validats per traductors i que, per tant, probablement necessitaran un nombre d'edicions més petit» (Martín Mor et al. 2016, 67).
    - *PE integrada en viu*: l'eina TAO està connectada constantment amb un sistema de TA extern i el posteditor rep resultats de la TA quan no hi ha cap segment que es puga aprofitar de la memòria de traducció. «Aquesta és la solució tecnològica que sembla tenir més projecció de futur, atés que els SGET cada cop faciliten més la interconnexió amb sistemes de TA externs» (Martín Mor et al. 2016, 67).
    - *PE integrada in vitro*: la TA sense posteditar es guarda en forma de memòria de traducció, sense que hi haja cap comunicació directa entre les eines. «Sia per motius de connectivitat, o fins i tot de seguretat, es pot decidir fer un projecte de PE sense donar accés directe al sistema de TA» (Martín Mor et al. 2016, 67).

Per acabar, pel que fa a la cronologia entre la publicació d'un text original i la traducció, la postedició permet obtenir una traducció final en un temps inferior. Tanmateix, hi ha casos en què es

publiquen les traduccions sense posteditar de certs productes, atés que les traduccions n'endarreixen l'aparició per a tota la comunitat. Aquest fenomen s'anomena *traducció desatesa* (Martín Mor et al. 2016, 68).

Tècniques de traducció automàtica

Pel que fa a les tècniques de traducció automàtica existents avui dia, hi ha tres sistemes principals: la traducció automàtica basada en regles lingüístiques (TABR o RBMT, de *rule-based machine translation*), la traducció basada en corpus (*corpus-based machine translation*) i la traducció automàtica neuronal (TAN o NMT, de *neural machine translation*).

La primera de les tècniques, la TABR, té una base totalment lingüística i n'hi ha tres tipus principals:

- **Sistemes directes:** es tracta de la traducció automàtica basada en diccionaris, paraula per paraula, amb regles molt bàsiques (Sindhu i Sagar 2016).
- **Sistemes de TABR de transferència:** són els sistemes de TABR més emprats avui dia. Es fonamenten en la transferència lèxica i sintàctica d'elements lingüístics d'una llengua a una altra per mitjà d'unes regles preestablertes: es «compilen diccionaris en forma electrònica, programen analitzadors morfològics i sintàctics, defineixen regles de transformació gramatical» (Ginestí i Forcada 2009) Els sistemes de TABR més coneguts són la plataforma lliure Apertium i Lucy Translation.
- **Sistemes de TABR basats en un llenguatge intermedi** (*interlingua*): la llengua de partida es transforma en un llenguatge intermedi (abstracte i independent), de manera que el text traduït es genera partint d'aquest llenguatge intermedi. Es tracta d'un sistema clàssic que va començar a idear-se a finals dels anys 50 (Richens 1958).

La segona tècnica, la traducció automàtica basada en corpus, es basa en mostres reals de textos, de traducció i textos monolingües per a generar traduccions automàtiques. Hi ha dos tipus principals:

- **Traducció automàtica estadística** (TAE o SMT, de *statistic machine translation*): amb aquest mètode s'entrenen motors de TA a partir de corpus de text bilingües i monolingües, de manera que aprenen a traduir aplicant models estadístics:

> «SMT is a machine translation paradigm that generates translations based on a probabilistic model of the translation process, the parameters of which are estimated from parallel text» (Yang i Min 2015, 201).
>
> Així doncs, no es basen en cap mena de coneixement lingüístic, sinó que es tracta d'un sistema purament estadístic. Un dels desavantatges és que es necessiten grans quantitats d'informació lingüística, emmagatzemada en corpus, els quals solen ser escassos en llengües minoritzades o amb un nombre baix de parlants. En aquesta categoria, hi destaca la plataforma de codi obert Moses i és el sistema al qual dedicaré el treball. A l'apartat següent (3.2. Traducció automàtica estadística) se n'explicarà el funcionament amb detall.

- **Traducció automàtica basada en exemples:** aquest sistema tradueix per analogia a altres traduccions trobades a corpus lingüístics, les quals funcionen com a exemples:

  > «When a new source sentence is to be translated, the examples are retrieved to find similar ones in the source side to match. Then the target sentence is generated by imitating the translation of the matched examples» (Qun i Xiaojun 2015, 112).

A més a més, hi ha un grup de motors de traducció, anomenats **híbrids** (HMT, de *hybrid machine translation*), que combinen els models estadístics amb les regles lingüístiques, com ara el Systran, la majoria de combinacions d'idiomes del traductor de Google, KantanMT o Bing Translator de Microsoft —a aquesta conclusió s'arriba per observació, atés que el codi dels traductors no és accessible—. La major part dels sistemes de traducció automàtica són híbrids en major o menor mesura:

> «RTMB may adopt a statistical word segmentation or parsing, while SMT usually utilizes human-encoded rules to translate certain types of names entities such as time, date, numerical expressions and name of persons, locations or organizations.
>
> Almost all the practical MT systems adopt hybrid approaches to a certain extent». (Qun i Xiaojun 2015, 115)

El sistema de traducció automàtica més innovador existent avui dia és el **neuronal** (TAN), el qual està basat en l'aprenentatge profund i pot traduir frases complexes d'una manera molt més fluida i amb una quantitat molt més baixa de dades, cosa que fa que siga un sistema molt prometedor en

el sector de la traducció, especialment pel que fa a la postedició: «for the automatic post-editing task, neural end to-end systems were found to represent a "significant step forward" over a basic statistical approach» (Castilho et al. 2017). Tanmateix, aquest sistema resulta car i també presenta inconvenients, com ara l'aparició de paraules estranyes:

> «Beyond being computationally expensive, it also often falters when it encounters unusual or rare words. And it's infamous for translating content only some of the time, resulting in weird, incomplete mixed-language translations.» (Hutchins 2017)

A més a més, hi ha qui defensa que actualment hi ha una certa exageració en relació amb la TAN i que els resultats no són tan clars:

> «NMT has generated great hype, especially as the translation industry is eager for improved MT quality in order to minimise costs. Although promising results are being reported when comparing NMT with other MT paradigms using automatic metrics, when human evaluation is added to the comparison, the results are not yet so clear-cut» (Castilho et al. 2017).

Google va estrenar, al novembre del 2016, aquest sistema per a les combinacions entre l'anglés i el francés, l'alemany, el castellà, el portugués, el xinés, el coreà i el turc (Turovsky 2016), i al març del 2017 va ampliar-lo al rus, l'hindi i el vietnamita (Turovsky 2017). D'altres proveïdors, com ara Prompsit, Amazon o Systran ja estan treballant en el desenvolupament d'aquest sistema de traducció automàtica (Koot 2016). D'altra banda, existeix un sistema de TAN de codi obert, anomenat OpenNMT[1], desenvolupat originalment per Yoon Kim i Harvard NPL i que avui utilitza i desenvolupa Systran.

3.2. Traducció automàtica estadística

La traducció automàtica estadística és un sistema de traducció automàtica que es basa en models estadístics probabilístics creats a partir de corpus monolingües i bilingües. Aquests models estadístics estan basats en la probabilitat, és a dir, en la versemblança que un esdeveniment donat tinga

1 URL: http://opennmt.net/

lloc. Dit altrament, aquest sistema parteix de traduccions humanes prèvies per a traduir de forma automàtica textos nous per mitjà de càlculs estadístics.

Els primers sistemes de traducció estadístics, com ara el motor de TAE d'IBM, llançat per primera vegada l'any 1999 pel Watson Research Center amb el projecte «Candide» (Chan 2015), prenien la paraula com a unitat bàsica de traducció. Tanmateix, aquests sistemes tenien moltes limitacions en tractar les paraules individualment, atés que una paraula en una llengua pot traduir-se per dues o més paraules en una altra, o viceversa, i que no es tenen en compte estructures lingüístiques de longitud major que la paraula. Així, els sistemes de traducció automàtica estadística més utilitzats avui dia, com el que s'estudia en aquest projecte, parteixen de segments bilingües de longitud variable, anomenats *phrases* en anglés.

Hi ha tres fases principals a la traducció automàtica estadística: l'entrenament (*training*), l'ajust dels paràmetres del sistema (*tuning*) i la traducció mateixa (*decoding*). A continuació, s'explicaran aquests tres processos principals de la traducció automàtica estadística.

3.2.1. Entrenament

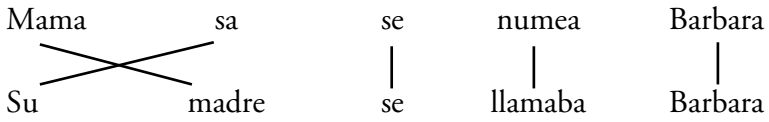
La fase d'entrenament comprén l'extracció de models estadístics de traducció a partir corpus lingüístics. L'objectiu principal és obtenir dos models estadístics: el model de la llengua de destí, extret de corpus monolingües (coneguts com a monotextos) i el model de traducció de segments, extret de corpus bilingües (també anomenats bitextos), és a dir textos paral·lels alineats en totes dues llengües. A més a més, hi ha d'altres models estadístics que s'extrauen durant la fase d'entrenament, com també d'altres paràmetres que s'empren com a complements dels dos models principals per tal d'obtenir una qualitat més alta. A continuació, s'explicaran amb més detall com funcionen tots aquests models i paràmetres.

3.2.1.1. Models estadístics i paràmetres de la TAE

Model de traducció de segments

El model de traducció $P(S|T)$[2] serveix per a calcular amb quina probabilitat un segment en la llengua d'arribada serà equivalent a un segment en la llengua de partida, és a dir, amb quina probabilitat es mantindrà el significat del segment de partida al segment d'arribada.

Aquest model estadístic té com a punt de partida un corpus bilingüe, és a dir, un conjunt de textos paral·lels alineats en totes dues llengües. L'objectiu és obtenir-ne un diccionari bilingüe probabilístic de segments de la següent manera: s'alineen les paraules de cada parell d'oracions per a identificar equivalències entre paraules. Aquesta alineació està basada en un model de traducció de paraules (Chan 2015), com en el cas dels motors d'IBM:



A continuació, aquest corpus bilingüe de paraules s'utilitza per a obtenir el diccionari de parells segments bilingües. Les paraules d'aquests parells de segments han de ser coherents en totes dues llengües (Koehn 2010), la qual cosa significa que no pot quedar cap paraula sense alinear ni al segment de partida ni al de destí. Un diccionari de parells segments podria ser el següent:

Partida (ro)	*Destí (es)*
mama	madre
sa	su
mama sa	su madre
se	se
numea	llamaba
se numea	se llamaba
mama sa se numea	su madre se llamaba
Barbara	Barbara

2 P=probabilitat, S=llengua de partida, T=llengua d'arribada.

numea Barbara	llamaba Barbara
se numea Barbara	se llamaba Barbara
mama sa se numea Barbara	su madre se llamaba Barbara

Finalment, es calculen les probabilitats de traducció de cada segment. A cada un dels segments s'assignaria un valor del 0 a l'1, en què l'1 és la probabilitat màxima.

El model de traducció sempre és *direccional* (Hearne i Way 2011), és a dir, funciona en una sola direcció i, a més a més, és un model «invers»: a partir d'un candidat en la llengua d'arribada (*T*), calcula amb quina probabilitat prové d'un cert segment d'arribada (*S*). Així, es necessiten dos models, un per a cada direcció de traducció, atés que la traducció no és simètrica. Per exemple, *dues* en català (o *două*, en romanés) sempre es tradueix per *dos* en castellà, mentre que el *dos* castellà es pot traduir al català tant com a *dos* i *dues* (o *doi* i *două* en romanés), en funció del gènere. Així, la probabilitat que *doi* es traduïsca per *dos* serà més alta [$P(dos|doi)=1$] que *dos* es traduïsca per *doi* [$P(doi|dos)=0,5$].

Model per a la ponderació lèxica

Aquest model parteix d'un diccionari bilingüe probabilístic de paraules —s'aprén a partir del corpus bilingüe probabilístic, també utilitzat en el model anterior per a establir la segmentació— i determina la fiabilitat d'un segment bilingüe. Koehn ho platenja així: «How can we judge if a rare phrase pair is reliable? If we descompose it into its word translations, we can check how well they match up» (Koehn 2010, 139). La raó d'emprar aquest model és que el model de traducció de segments anterior és poc fiable en el cas de segments més llargs i, per tant, infreqüents: encara que els parells llargs obtinguen una probabilitat alta de traducció, com que no és comú que apareguen moltes vegades, la fiabilitat és més baixa, atés que, com més llarg siga un segment, menys vegades apareixerà en el corpus.

Així doncs, aquest model es basa en un diccionari probabilístic bilingüe obtingut a partir d'alineacions de paraules (no de segments). El que fa exactament és cercar als segments les traduccions paraula per paraula per veure'n la coincidència amb el diccionari probabilístic bilingüe. Si la puntuació de la probabilitat a escala de paraula és major, els resultats seran més fiables. La fórmula per la ponderació lèxica de l'oració romanesa «șeful vine mâine» quedaria així:

$$lex(el\ jefe\ viene\ mañana|ședul\ vine\ mâine) = w(\frac{el|ședul + jefe|ședul}{2}) \times w(viene|vine) \times w(mañana|mâine)$$

S'hi multipliquen les probabilitats de cada una de les paraules en la llengua de destí. Si hi ha més d'una paraula d'origen per a la paraula de destí, es divideix pel nombre de total de paraules d'origen.

Model de la llengua d'arribada

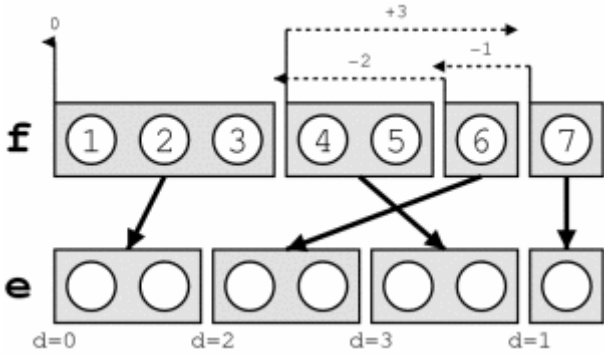
Aquest model serveix per a estimar la qualitat i la fluïdesa de la traducció en la llengua d'arribada i parteix de corpus monolingües en aquesta llengua. El model està basat en *n-grames* (segments de *n* paraules) i s'entrena a partir de grans corpus monolingües de textos en la llengua d'arribada. Dit d'una altra manera, compara la traducció en la llengua d'arribada amb segments de corpus monolingües per determinar si la traducció és fluida i natural o no. Es pot calcular fàcilment analitzant la freqüència dels *n-grames* trobats durant l'entrenament aquest model estadístic. S'empren *n-grames* perquè és molt difícil que s'haja trobat abans una oració completa. El nombre (*n*) d'*n-grames* pot ser variable, tot i que el més emprat és el de 5-grames: «the n-gram model that is normally used by SMT programs is the 5-gram model» (Pérez García 2016).

Aquest seria un exemple del funcionament d'un model de traducció basat en trigrames (*3-grames*) per a l'oració «Long fue pionero en el uso de la antestesia»:

p(Long fue pionero en el uso de la anestesia) = *p*(Long) x *p*(fue|Long) x *p*(pionero|Long fue) x *p*(en|fue pionero) x *p*(el|pionero en) x *p*(uso|en el) x *p*(de|el uso) x *p*(la|uso de) x *p*(anestesia|de la)

Model de reordenament basat en distàncies

Aquest model serveix per a tenir en compte el reordenament basat en la distància. Es compta el nombre de paraules omeses quan els segments del text d'arribada no estan en el mateix lloc que els de la llengua d'origen, és a dir, quant ha variat la posició dels segments.



Il·lustració 1: esquema del model de reordenament basat en distàncies extret del manual de Moses (Moses 2017)

Model de reordenament lexicalitzat

Aquest model de reordenament parteix de les alineacions obtingudes en el procés d'extracció de segments bilingües. Serveix per a mesurar l'ordre en què les paraules alineades apareixen en el text de destinació. Per al reordenament també es recorre a aquest segon mètode perquè l'anterior no és prou sòlid, atés que no té en compte paràmetres lèxics:

> «However, the distance-based model is fairly weak and, consequently, there is a need to introduce another model that conditions the reordering on the specific phrases that are being reordered» (Pérez García 2016, 17).

Es poden donar tres situacions pel que fa al reordenament:

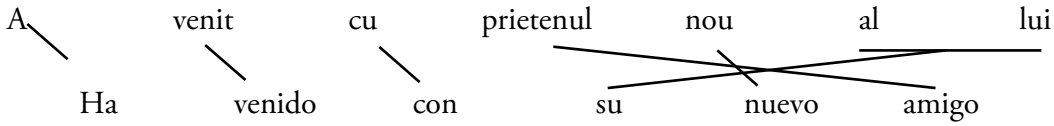
→ *Monotone:* no es dóna el reordenament.

Maria	vine	mâine	din	București
Maria	viene	mañana	de	Bucarest

→ *Swap:* el segment d'arribada canvia la seua posició amb la traducció de l'anterior.

Mama	sa	se	numea	Barbara
Su	madre	se	llamaba	Barbara

→ *Discontinous:* cap de les anteriors.



Aquest model funciona de manera que divideix el nombre de vegades que s'ha dut a terme una operació de reordenament pel nombre de vegades que s'ha extret el parell de segments del corpus.

Penalització pel nombre de paraules

És un paràmetre que té en compte el nombre de paraules de la traducció en llengua d'arribada. S'hi recorre per tal d'evitar traduccions molt curtes, atés que el model de llengua prefereix sempre les traduccions més curtes, com explica Koehn: «Yet one complement of the system prefers shorter translations: the language model, simply because fewer trigrams have to be scored» (2010, 140). D'altra banda, les traduccions amb més paraules solen tenir puntuacions més baixes.

Penalització pel nombre de segments bilingües

Aquest altre paràmetre té en compte el nombre de segments bilingües emprats per a generar la traducció en llengua de destí. S'hi recorre per a fomentar l'ús de segments llargs en funció de la finalitat del projecte, com explica Pérez García: «Depending of the scope of the project, different phrase lengths may be needed» (2016, 19).

3.2.1.2. Combinació dels models i ajustament dels paràmetres del sistema

Com que no tots els models anteriors tenen la mateixa importància, s'assigna un pes diferent a cadascun dels models per tal de trobar un valor de la probabilitat final de la traducció. Per tal de trobar la màxima qualitat de les traduccions, s'ajusten els paràmetres del sistema (*tuning*) a partir de valors aleatoris dels pesos de cada model. L'ajustament dels paràmetres del sistema també s'anomena MERT, de les sigles en anglés per a *minimum error rate traning (Koehn 2010, 264).*

Per a dur a terme aquest ajustament, cal poder avaluar la qualitat de les traduccions de manera automàtica a partir d'un corpus paral·lel diferent de l'emprat per a entrenar els models. Aquests mètodes d'avaluació automàtica de la traducció també serveixen per a analitzar la qualitat dels resultats del motor de traducció automàtica i se'n parlarà amb més profunditat a l'apartat 3.2.3.

3.2.1.3. Ús de recursos especialitzats

Els motors de TAE se solen especialitzar en dominis concrets per aconseguir una qualitat de la traducció més alta. Això es fa mitjançant l'entrenament dels motors a partir de corpus especialitzats en la temàtica corresponent. Aquests corpus es poden complementar amb d'altres de dominis no relacionats amb l'especialitat del traductor: «It is frequently necessary to supplement scarce in-domain training data with out-of-domain data» (Haddow i Koehn 2012, 1).

Tanmateix, l'ús de corpus no pertanyents al domini del motor, també pot tenir conseqüències negatives:

> «For example, a large out-of-domain data set may improve the translation of rare words by providing better coverage, but degrade translation of more common words by providing erroneous out-of-domain translations» (Haddow i Koehn 2012, 2).

Al mateix estudi de Haddow i Koehn s'analitza la qualitat de la TA comparant motors entrenats amb corpus d'un sol domini i un d'entrenat amb corpus pertanyents i no pertanyents al domini. El resultat del motor amb corpus mixts no és superior a un dels motors entrenat només amb corpus especialitzats. També s'hi mostra que entrenar el motor amb corpus no pertanyents al domini és útil per a disminuir la quantitat de paraules desconegudes, i es conclou que caldria assignar pesos diferents a cada corpus:

> «This explains why approaches which try to weight the out-of-domain data in some way (e.g. corpus weighting or instance weighting) can be more successful than simply concatenating data sets» (Haddow i Koehn 2012, 9-10).

3.2.1.4. Preparació dels textos per a l'entrenament del motor de TAE

Els textos emprats per a entrenar el motor de TAE s'han de processar abans de l'entrenament. Aquesta preparació de textos consta de tres passos bàsics i es duen a terme de manera automàtica gràcies als diversos components del programa:

- Neteja (*cleaning*) del corpus: el corpus es neteja per tal que les dades que s'introdueixen al motor siguen de prou qualitat. Aquest pas inclou assegurar que els formats de fitxer i la codificació de caràcters és l'adient, eliminar duplicats, eliminar segments idèntics a l'ori-

gen i a la llengua de destí, etc. Koehn indica que pot ser recomanable dur a terme aquesta neteja, però que els motors de TAE haurien de ser prou potents per a menystenir les dades de baixa qualitat: «Statistical machine translation models are generally assumed to be fairly robust to noisy data, such as data that includes misalignments» (2010, 61).

- Tokenització (*tokenization*): s'afigen espais entre les paraules i els signes de puntuació.

```
Tenga especial cuidado con Kinzalkomb ; informe a su médico
si padece o ha padecido alguno de los siguientes trastornos o
enfermedades :
```

- Conversió de majúscules i minúscules (*truecasing*): conversió de les paraules inicials de cada oració a la seua forma més freqüent. Els noms propis es queden en majúscules.

```
tenga especial cuidado con Kinzalkomb ; informe a su médico si
padece o ha padecido alguno de los siguientes trastornos o
enfermedades :
```

Després d'aquestes tres passes, ja pot començar l'entrenament del motor de TAE. Aquest procés pot ser llarg, en funció de la quantitat de dades que s'hi hagen introduït. La part pràctica de l'entrenament d'un motor de TAE, en aquest cas a la plataforma MTradumàtica, està explicada amb detall a l'apartat 4.3.

3.2.2. Traducció

Una vegada s'ha entrenat el motor de TAE, ja es pot utilitzar per a traduir textos. A continuació, es descriuen les passes que segueix el motor per tal de traduir un text.

Preparació dels textos

Tal com ocorre amb els textos emprats per a entrenar el motor de TAE, els textos també s'han de processar prèviament per tal que el motor els puga traduir. En aquest cas, hi ha dos processos per aplicar als textos per traduir: la tokenització i la conversió de majúscules a minúscules, tots dos explicats a l'apartat 3.2.1.4.

Traducció

Per dur a terme la traducció (*decoding*), quan introduïm una oració en llengua de partida per traduir se cerquen totes les traduccions possibles, s'assigna una puntuació a cada model, es calcula la puntuació final i es tria la traducció amb la puntuació més alta.

Això implica que s'haurien d'avaluar totes les maneres en què es podria segmentar l'oració, traduir-la de totes les maneres possibles i provar tots els reordenaments possibles. Tanmateix no hi ha cap garantia que es puga trobar la millor traducció, atés que això no és possible i s'empren regles heurístiques per tal de no generar totes les traduccions possibles.

Així, en cercar la traducció d'un segment d'origen, es recullen alternatives de traducció per trobar el millor camí de traducció: es calcula la puntuació parcial i s'estima la puntuació general per generar hipòtesis parcials de traducció, es rebutgen aquelles que són poc esperançadores i es recombinen les hipòtesis parcials semblants.

Un dels problemes que es poden trobar durant la segmentació de l'oració és el fet que hi apareguen paraules desconegudes. En aquest cas, és possible que el motor no segmente les oracions correctament, cosa que provoca que el resultat siga molt pitjor. Una mala segmentació afecta també el reordenament, la selecció lèxica i, en conseqüència, el significat general de l'oració.

Postprocessament dels textos traduïts

Una vegada s'ha trobat la traducció final, s'ha de dur a terme el procés invers al de la preparació dels textos per traduir. Així, és necessari:

- Desfer la conversió de majúscules i minúscules (*truecasing*):

```
Aveţi grijă deosebită când utilizaţi Kinzalkomb ; vă rugăm să spuneţi
medicului dumneavoastră dacă suferiţi sau aţi suferit de vreuna
dintre următoarele afecţiuni sau tulburări :
```

- Destokenització: s'esborren els espais entre les paraules i els signes de puntuació.

```
Aveţi grijă deosebită când utilizaţi Kinzalkomb; vă rugăm să
spuneţi medicului dumneavoastră dacă suferiţi sau aţi suferit
de vreuna dintre următoarele afecţiuni sau tulburări:
```

3.2.3. Avaluació automàtica de la qualitat de la traducció automàtica

Com s'ha explicat anteriorment a l'apartat 3.2.1.2., en la TAE és important avaluar els resultats de la traducció automàtica amb dos objectius: poder modificar els paràmetres del sistema (*tuning*) i avaluar la qualitat mateixa dels resultats de traducció del motor entrenat. Per a dur a terme aquesta avaluació, cal reservar una part del corpus bilingüe, la qual no s'emprarà per a l'entrenament del motor, com a referència per als sistemes d'avaluació automàtica de la qualitat. Aquest extracte del corpus se sol anomenar *gold standard* (Dorr et al. 2011, 853). Avui dia s'empren diversos mètodes d'avaluació automàtica de la traducció automàtica; se n'exposaran breument, sense entrar en detalls, tres dels més emprats: BLEU, METEOR i TER.

BLEU

El mètode BLEU (Papineni et al. 2002)—de les sigles en anglés *Bilingual Evaluation Understudy* — és un dels més famosos per a l'avaluació automàtica de la TA. Té en compte l'ordre de les paraules de manera parcial dins dels segments, partint del resultat de la TA i d'una o més traduccions humanes de referència:

> «It is based on counting the number of n-grams, namely sequences of consecutive word(s) of varying length, co-occurring in an MT output and in one or more versions of corresponding reference, usually each in the form of a sentence». (Chunyu i Tak-ming 2015, 226).

Aquest mètode puntua les traduccions amb una nota del 0 a l'1, sent l'1 la qualitat òptima. Es considera que la qualitat del motor de TA és bona quan la puntuació és superior del 0,5:

> «A score greater than 50% is a very good score and significantly less post-editing will be required to achieve publishable translation quality» (KantanMT 2017).

Tot i això, té un inconvenient i és que no té en compte les coincidències aproximades, com ara la traducció amb sinònims o estructures equivalents.

METEOR

El mètode METEOR (Banerjee i Lavie 2005) es basa en les alineacions paraula a paraula entre els textos produïts pel motor de TA i les traduccions humanes de referència. Un dels seus objectius és millorar el sisitema BLEU i incloure-hi paràmetres que no estenien en compte. Ho expliquen Chunyu i Tak-ming:

> «To maximize the possibility of matching, it uses three word-mapping criteria: (1) exact character sequences, (2) identical stem forms of word, and (3) synonyms» (2015, 228).

Així, a diferència de BLEU, està basat en la precisió (la proporció de paraules a la TA que apareixen a la referència) i en la cobertura o *recall* (proporció de paraules de la referència que apareixen a la traducció automàtica).

TER

El mètode TER (Snover et al. 2006)—de les sigles en anglés de *Translation Edit Rate*— mesura l'esforç requerit perquè la postedició del text siga de la mateixa qualitat que la traducció humana:

> «TER measures the amount of editing that a human would have to perform to change a system output so it exactly matches a reference translation» (Snover et al. 2006)

Quantifica la distància d'edició entre dues oracions, és a dir, en el nombre mínim d'operacions necessàries per a transformar una oració en una altra. És un indicador negatiu, per la qual cosa 0 és el valor òptim.

3.3. MTradumàtica

La plataforma emprada per a l'entrenament dels motors de traducció automàtica estadística és MTradumàtica, la qual forma part del grup de recerca ProjecTA[3], pertanyent a Tradumàtica [4] del Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental de la Universitat Autò-

3 URL: http://www.projecta.tradumatica.net/

4 URL: http://www.tradumatica.net/

noma de Barcelona. L'objectiu del grup de recerca és acostar la TA als traductors i a les empreses de traducció, com explica Martín-Mor:

> «It works from the basic assumption that translators have the appropriate profile to manage MT-related tasks, and that empowering translators in MT tasks is beneficial for translation companies» (Martín-Mor 2017, 1).

Aquesta eina de personalització de motors de traducció automàtica estadística està basada en Moses,[5] un dels sistemes de personalització més coneguts, lliure, gratuït i amb llicència LGPL, i ha estat desenvolupada per l'empresa Prompsit.[6]

Actualment, MTradumàtica està disponible a la xarxa com a plataforma web,[7] encara que també es pot instal·lar a ordinadors i a servidors:

> «It is currently available for testing purposes and can be installed stand alone on a PC as well as on a server so that it can be accessed from any computer with an internet connection via a web browser» (Martín-Mor 2017, 7)

La plataforma permet entrenar motors de traducció automàtica estadística personalitzats, de manera intuïtiva, gràfica i clara, per tal d'entendre el funcionament de cada procés (des de la pujada de fitxers i la creació de corpus monolingües i bilingües, fins a l'entrenament del model de llengua i l'entrenament mateix del model de traducció i del motor). El funcionament d'aquests processos s'explica de manera pràctica i amb detall a l'apartat 4.3. A banda d'això, gràcies la funció «Inspect», es pot analitzar el funcionament dels models per tal de comprendre millor en què consisteix el procés de TAE.

MTradumàtica continua en desenvolupament, de manera que oferirà funcionalitats molt interessants, com ara la pujada de fitxers en format TMX (actualment només accepta text pla), l'analisi del model de traducció, la gestió d'usuaris, la priorització de models, la gestió de terminologia i de corpus. A més a més, permetrà la preedició automàtica i la postedició de textos, l'avaluació auto-

5 URL: http://www.statmt.org/moses/

6 URL: http://www.prompsit.com/inici/

7 URL: http://m.tradumatica.net/

màtica de la qualitat de la TA i la integració amb eines TAO per mitjà d'API, com explica Martín-Mor (2017, 14-15).

3.4. El romanés i la TA

3.4.1. Particularitats morfològiques que poden tenir impacte en la TAE

D'acord amb la *Gran Enciclopèdia Catalana*, el romanés és una «llengua romànica del grup oriental, parlada per més de vint milions de persones, a Romania, a Moldàvia i, en graus diversos, a tots els països de la península Balcànica (on constitueix nombroses illes dialectals, disseminades al territori de Grècia, Croàcia, Sèrbia, Montenegro, Macedònia, Albània i Bulgària)» (Gran enciclopèdia catalana 2012). L'edició web de *Ethnologue* compta gairebé 24 milions de parlants de romanés a tot el món (Ethnologue: Languages of the World, Seventeenth edition 2013), per la qual cosa és la 36a llengua per nombre de parlants i la desena d'Europa.

Tot i que l'origen del romanés és el mateix que el de la resta de llengües romàniques, la fragmentació del llatí, la llengua romanesa presenta particularitats úniques que no comparteix amb la resta de llengües romàniques. Moltes d'aquestes particularitats estan relacionades amb la morfologia del romanés i poden tenir un efecte en els resultats d'un motor de TAE.

Per començar, el romanés és l'única llengua romànica que continua declinant substantius, adjectius, determinants i pronoms, és a dir, modifica el final de les paraules d'acord amb la funció gramatical. A més a més, continua conservant el gènere neutre, el qual funciona com el masculí en singular i com a femení en plural. D'altra banda, la formació dels plurals és molt irregular i no té per què obeir a cap regla.

Un altre aspecte de la morfologia destacable és l'existència de quatre tipus d'article (definit, indefinit, possessiu-genitiu i demostratiu-adjectival) (Daniliuc i Daniliuc 2000). A continuació, per tal d'il·lustrar aquests fenòmens en donaré uns quants exemples, juntament amb la traducció al català i al castellà.

Un dels articles que pot interferir més en la traducció automàtica estadística és l'article definit. La raó és que, a diferència de la resta de llengües romàniques, l'article definit es construeix com a un sufix del substantiu mateix, per la qual cosa formen una única paraula. Vejam-ne uns quants

exemples: *cap/cap**ul*** (cap / el cap, *cabeza / la cabeza*), casă/cas**a** (casa / la casa), *pacienți/pacienț**ii*** (pacients / els pacients, *paciente / los pacientes*), *studii/studii**le*** (estudis / els estudis, *estudios/ los estudios*). Com hem dit anteriorment, aquests substantius articulats s'han de declinar per a crear diverses estructures, com ara el complement del nom. Així, *cap**ul** pacient**ului*** vol dir «el cap del pacient», amb *pacient* en genitiu i articulat.

Pel que fa a l'article indefinit, tot i que funciona de manera semblant al castellà o al català, construït davant del nom com a paraula independent, també es pot declinar (p. ex.: *înlocuirea **unor** pacienți* significa «la substitució d'uns pacients» / *la sustitución de unos pacientes*).

Els altres dos articles són més complexos i s'utilitzen en situacions més concretes, tot i que són molt usuals. Com a exemple de construcció amb article possessiu-genitiu podríem indicar «*Un prieten **al** Mariei*» («Un amic de Maria» / «*Un amigo de María*»). D'altra banda, un exemple d'oració amb article demostratiu adjetival podria ser: «***Cea** mai mare doză*» («la dosi més gran» / «*la dosis más grande*»).

Aquesta petita explicació, en què m'he centrat en l'article com a exemple, no té la intenció de fer una introducció a la llengua romanesa, atés que, entre altres coses, el motor estadístic no té en compte cap coneixement lingüístic. La raó de dedicar-hi un capítol és, doncs, de destacar que la flexibilitat de la llengua romanesa i la diversitat de la morfologia poden afectar els càlculs probabilístics del motor. Un clar exemple n'és l'articulació posterior dels substantius, fet que no ocorre en castellà. Com a exemple, en castellà, el motor reconeixerà la paraula *cabeza* tant al grup «*la cabeza*» com a «*una cabeza*». En canvi, en romanés, s'empren dues paraules que pot considerar diferents (*un cap* / *capul*).

3.4.2. Escriptura del romanés a dispositius informàtics

Tot i que en altres períodes el romanés s'ha escrit en l'alfabet ciríl·lic, l'alfabet romanés actual està basat en el llatí, i té les lletres següents:

A, Ă, Â, B, C, D, E, F, G, H, I, Î, J, K, L, M, N, O, P, R, S, Ș, T, Ț, U, V, X, Z.

Com s'observa, hi ha cinc lletres diacrítiques (Ă, Â, Î, Ș, Ț), l'escriptura de les quals no sempre ha estat fàcil als dispositius informàtics, en especial pel que fa a les lletres Ș i Ț (essa i te amb coma).

Aquestes dues lletres (quatre caràcters, tenint en compte les majúscules i les minúscules) no apareixien a les primeres versions d'Unicode, i no s'hi van incloure fins al setembre del 1999, amb la versió 3.0.0 d'Unicode (Unicode Consortium 2000). Unicode hi descriu el caràcter Ș com a «latin capital letter s with comma below» (U+0218), ș com a «latin small letter s with comma below» (U+0219), Ț com a «latin capital letter t with comma below» (U+021A), ț «latin small letter t with comma below» (U+021B). Un any més tard, ISO publica la ISO/IEC 8859-16 amb els caràcters correctes per al romanés.

Fins aleshores, els caràcters utilitzats eren Ş i Ţ, emprats al turc i disponibles a Unicode des de la versió 1.1.0, del juny del 1993. A la versió del 1995, el Consorci Unicode estableix que els caràcters que s'apliquen tant al romanés com al turc són Ş (latin capital letter s with cedilla: U+015E), ş (latin small letter s with cedilla: U+015F), Ţ (latin capital letter t with cedilla: U+0162), ţ (latin small letter t with cedilla: U+0163). «Turkish language indeed uses cedilla in U+015E, U+015F but does not make any use of U+0162, U+0163. Romanian language doesn't use any of them» (Kaplan 2011).

Aquest fenomen ha provocat que durant tota la dècada passada s'hagen donat problemes de compatibilitat tant als diferents sistemes operatius i que, a banda de no existir-hi els caràcters correctes, en molts casos era difícil instal·lar-hi el teclat. Kaplan fa una bona descripció sobre aquests problemes de compatibilitat de manera detallada (Kaplan 2011).

El fet que la instal·lació del teclat i de la configuració de l'idioma als ordinadors siga tan complicada ha fet que molts usuaris de romanés opten per escriure sense diacrítics o amb els incorrectes. Com a resultat, gran part dels textos informatitzats en romanés estan escrits amb els caràcters turcs o directament sense cap diacrític.

4. Metodologia

L'objectiu d'aquest apartat és d'exposar les passes que s'ha seguit per a entrenar un motor de traducció automàtica estadística des d'un punt de vista pràctic. S'hi exposarà cadascun dels procediments tècnics a què s'ha hagut de recórrer en cada una de les etapes: la cerca i descàrrega de recursos lingüístics, el processament de textos i la conversió de formats, l'entrenament mateix dels motors a la plataforma MTradumàtica i, finalment, la metodologia per a l'avaluació automàtica de la qualitat de la traducció automàtica[8].

És important recordar que, pel funcionament dels motors de traducció estadística, el fet que el motor haja de ser bidireccional (es→ro i ro→es) implica que, en realitat, s'han d'entrenar dos motors diferents per separat, un per a cada combinació d'idiomes. Pel que fa als recursos que s'hi utilitzen, el model de traducció es pot entrenar amb el mateix corpus paral·lel bilingüe, encara que cada model de llengua de destí es pot entrenar amb corpus monolingües d'altres procedències.

4.1. Cerca de recursos lingüístics

La primera etapa per a l'entrenament d'un motor de traducció automàtica estadística és identificar i obtenir recursos lingüístics amb què entrenar-lo.

L'èxit en l'obtenció dels recursos depén, principalment, de la combinació de llengües del motor de TAE, atés que la xarxa ofereix més possibilitats i recursos per a les llengües amb més usuaris. En els meus motors de traducció automàtica estadística entren en joc dues llengües: una de majoritària i amb gran quantitat de recursos (el castellà) i una amb menys recursos a la xarxa (el romanés). A més a més, com que es tracta d'un motor de traducció estadística especialitzat, la dificultat pot ser més gran, atés que els corpus també haurien de pertànyer al domini. En el meu cas, els materials necessaris per models de traducció i la dificultat teòrica per a trobar-los podrien ser els següents:

8 A l'Annex 1 s'inclou un guia en què s'explica pas a pas, de manera resumida i concreta, com s'ha entrenat el motor a la plataforma MTradumàtica. També s'hi troben enllaços als recursos creats durant el procés.

→ Motor de traducció ro→es:

- Per al model de traducció: corpus bilingües ro↔es sobre medicina i farmàcia. Pot resultat complicat pel domini d'especialització i perquè és necessari que siguen corpus paral·lels.
- Per al model de llengua de destí (es): corpus monolingüe sobre medicina i farmàcia. Pot ser més fàcil de trobar-ne, perquè hi ha més recursos en castellà. A més a més, es poden utilitzar també els textos en castellà del corpus bilingüe.

→ Motor de traducció es→ro:

- Per al model de traducció: corpus bilingües ro↔es sobre medicina i farmàcia. Empraré el mateix que per a l'altre motor.
- Per al model de llengua de destí (ro): corpus monolingüe sobre medicina i farmàcia. La quantitat de recursos en romanés és inferior que la de castellà, fet que pot portar dificultats. També es poden utilitzar els textos en romanés del corpus bilingüe.

Així, per resumir, es necessiten:

- Corpus bilingües sobre medicina i farmàcia ro↔es per a tots dos models de traducció i tots dos models de llengua de destí.
- Corpus monolingüe en castellà per a millorar el model de llengua de destí (es).
- Corpus monolingüe en romanés per a millorar el model de llengua de destí (ro).

Com s'observa, i relacionat amb el que s'exposa a l'apartat 3.2.1.3, en aquest cas no s'ha seleccionat cap corpus no pertanyent al domini per a l'entrenament. A continuació, s'exposaran els recursos trobats i la seua procedència.

4.1.1. OPUS Corpus: EMEA

El primer lloc que es va visitar ser el lloc web del projecte OPUS,[9] el qual s'encarrega, des de començaments de l'any 2000, de col·leccionar textos paral·lels i oferir-los de manera gratuïta. L'any 2012 la col·lecció «comptava 90 idiomes i incloïa informació de diversos dominis» (Tiedemann 2012). En seleccionar la combinació es>ro, els resultats van ser els següents:

9 URL: http://opus.lingfil.uu.se/

Search & download resources: es (Spanish) ro (Romanian) all

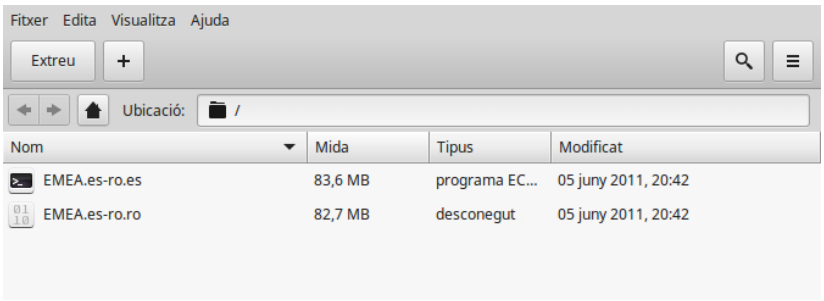
Language resources: click on [tmx | moses | xces | lang-id] to download the data! (raw = untokenized, true = truecaser model, TM = phrase-based translation model)

corpus	doc's	sent's	src tokens	trg tokens	XCES/XML	raw	TMX	Moses	mono	raw	true	TM	dic	freq		Browse Files
OpenSubtitles2016	45437	37.3M	280.2M	272.3M	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro					[sample]	[xml/es][xml/ro][raw/es][raw/ro]
OpenSubtitles2013	24683	21.8M	166.4M	166.0M	[xces es ro]	[es ro]		[moses]		es ro			dic	es ro	[sample]	
OpenSubtitles2012	24248	21.3M	160.6M	161.4M	[xces es ro]	[es ro]		[moses]		es ro	es ro	es-ro	dic	es ro	[sample]	
DGT	12095	1.6M	34.7M	33.0M	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro	es ro	es-ro		es ro	[sample]	[xml/es][xml/ro]
JRC-Acquis	6499	0.5M	22.2M	23.7M	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro					[sample]	[xml/es][xml/ro]
EMEA	1627	1.0M	12.6M	15.0M	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro	es ro	es-ro	dic	es ro	[query]	[sample] [xml/es][xml/ro]
EUbookshop	716	0.3M	13.1M	12.9M	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro	es ro	es-ro	dic	es ro		[sample] [xml/es][xml/ro]
Europarl	6559	0.4M	11.1M	10.7M	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro	es ro	es-ro	dic	es ro	[query]	[sample] [xml/es][xml/ro]
GNOME	1275	0.7M	3.0M	3.1M	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro						[sample] [xml/es][xml/ro]
OpenSubtitles	259	0.3M	2.2M	2.2M	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro	es ro	es-ro	dic	es ro	[query]	[sample] [xml/es][xml/ro][xml/es-ro]
Tatoeba	1	0.2k	1.7M	38.7k	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro	es ro	es-ro	dic	es ro		[sample] [xml/es][xml/ro]
KDE4	1064	0.1M	1.1M	0.6M	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro	es ro	es-ro	dic	es ro	[query]	[sample] [xml/es][xml/ro]
Ubuntu	407	0.1M	0.5M	0.4M	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro						[sample] [xml/es][xml/ro]
Tanzil	1	11.0k	0.2M	0.2M	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro	es ro	es-ro		es ro	[query]	[sample] [xml/es][xml/ro][xml/es-ro]
PHP	3217	32.9k	0.2M	91.7k	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro	es ro	es-ro	dic	es ro	[query]	[sample] [xml/es][xml/ro]
GlobalVoices	74	2.5k	61.1k	58.3k	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro	es ro	es-ro	dic			[sample] [xml/es][xml/ro]
KDEdoc	6	0.2k	3.9k	2.2k	[xces es ro]	[es ro]	[tmx]	[moses]	es ro	es ro	es ro	es-ro	dic	es ro	[query]	[sample] [xml/es][xml/ro]
total	128168	85.4M	709.9M	701.5M	85.4M		62.5M	78.9M								

Il·lustració 2: recursos disponibles a l'OPUS en castellà i romanés.

Després de donar una ullada als tipus de corpus, el més adient va ser descarregar el de la EMEA, amb 12,6 milions de paraules en castellà i 15 milions en romanés i un milió de frases alineades. L'EMEA és l'Agència Europea de Medicina i el corpus naix a partir de l'alineació de documents en PDF d'aquesta organització, per la qual cosa el contingut s'adiu al camp d'especialització del motor. En aquest portal, els recursos lingüístics es poden descarregar en diversos formats (tmx, moses, text pla, xml, xces, etc.). En el meu cas, només cal descarregar la versió Moses, atés que és el sistema que empra MTradumàtica.

En fer clic a «moses» descarreguem una carpeta zip anomenada «es-ro.txt.zip», que haurem d'extraure. A dins trobarem dos fixters de text, cadascú amb el codi iso de la llengua corresponent com a extensió del fitxer. En algunes carpetes zip «Moses» de l'OPUS, també hi apareix un fitxer ids, en què hi ha informació sobre l'alineació del corpus. Aquests fitxers ja es podrien pujar directament a MTradumàtica, per la qual cosa, no és necessària la conversió dels fixters a cap altre for-

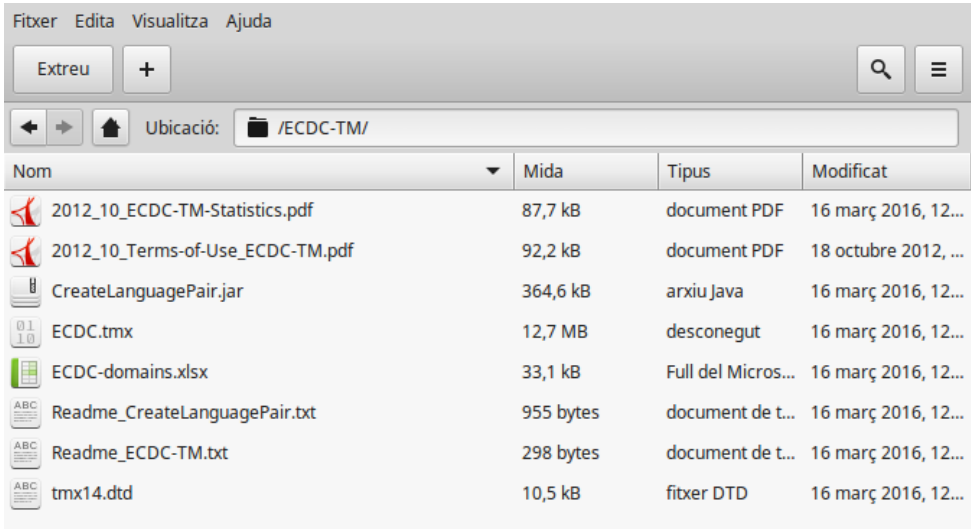


Il·lustració 3: contingut de la carpeta ZIP amb els fitxers en format Moses.

mat.

4.1.2. ECDC

El portal EU Science Hub,[10] el servei de ciència i coneixement de la Comissió Europea, també dedica un dels seus apartats a recursos de lingüístics tecnològics. S'hi inclouen recursos molt interessants, com ara totes les memòries de traducció des del 2004 de la Direcció General de Traducció (DGT) o les del Centre Europeu per a la Prevenció i el Control de les Malalties (ECDC,[11] per les sigles en anglés), les quals eren adients en el meu cas. Per a descarregar-la, només cal fer clic a «Download the ECDC Translation Memory» i fer clic a l'enllaç a la carpeta zip. En aquest cas, el portal no ofereix cap varietat d'opcions pel que fa al format dels fitxers, sinó que facilita una memòria de traducció en format tmx amb tots els idiomes. En descarregar la carpeta zip, s'observa que el contingut és molt diferent del cas anterior:



Il·lustració 4: contingut de la carpeta zip d'ECDC.

Efectivament, s'hi inclou un fitxer tmx acompanyat d'un programa en format Java per tal d'extraure els parells d'idiomes. A l'apartat 4.2.2 s'explica el procés de conversió i extracció del corpus.

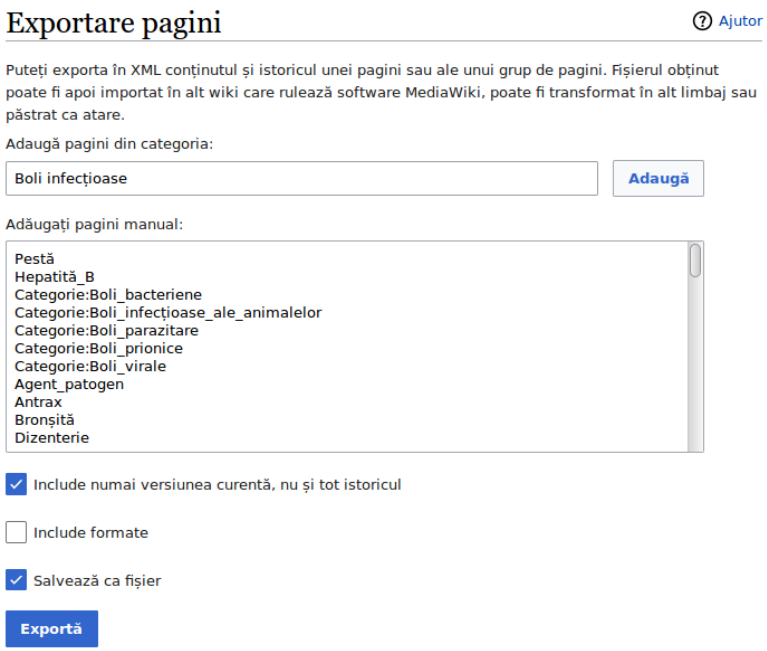
10 URL: https://ec.europa.eu/jrc/en/language-technologies

11 URL: https://ec.europa.eu/jrc/en/language-technologies/ecdc-translation-memory

4.1.3. Viquipèdia

Com que el meu objectiu era obtenir més recursos lingüístics especialitzats en medicina i farmàcia per als models de llengua, un bon lloc on trobar-ne, tant en castellà com en romanés podria ser la Viquipèdia. L'ús d'aquesta enciclopèdia lliure pot tenir molt de potencial, atés que, no sols hi ha una gran quantitat d'informació, sinó perquè els articles estan classificats per categories. Aquesta característica podria facilitar la descàrrega de recursos per dominis.

La Viquipèdia ofereix la funcionalitat d'exportar llocs web per categories i pàgines, gràcies a la funció especial exporta, accessible des d'aquí.[12] Així, afegint al cercador categories que ens interessen, l'eina afegeix a la llista tots els articles pertanyents a aquestes categories. Podem afegir totes les categories que vulguem, per la qual cosa la feina més feixuga pot ser de cercar categories adients. En el següent exemple s'observa com la Viquipèdia ha afegit tots els articles de la categoria «Boli Infecțioase» (malalties infeccioses) a la llista:



Il·lustració 5: funció d'exportació de Viquipèdia per categories.

Una vegada afegides totes les categories rellevants, podem fer clic a «Exportă» i Viquipèdia generarà un fitxer xml tots els articles de cada categoria. Pel que fa a les tres opcions d'exportació, el més adient és seleccionar la primera (inclou només la versió actual, sense l'historial complet) i la

12 URL: https://ro.wikipedia.org/wiki/Special:Export%C4%83

tercera (Desa com a fitxer). La segona no resulta interessant en aquest cas, atés que serveix per a incloure-hi el format, mentre que l'objectiu és obtenir-ne text pla. El tractament i l'extracció de text del fitxer xml es tractarà à a l'apartat següent.

Per descarregar els fitxers s'ha elaborat un llistat de categories tot navegant pels portals de medicina i farmàcia de la Viquipèdia. Finalment, s'han descarregat 249 articles en romanés i 429 articles en castellà.

4.2. Processament de textos i conversió de formats

Aquest apartat tracta sobre els mètodes seguits per a processar els textos descarregats i convertir-los en un format adient per a MTràdumàtica. Recordem que l'objectiu és obtenir un fitxer de text pla per a cada llengua amb «.es» o «.ro» com a extensió, segons el cas.

Per al tractament del corpus de l'EMEA no ha sigut necessari fer cap processament dels fitxers, atés que el web de l'OPUS ofereix l'opció de descarregar-los en un format compatible per a Moses. Tot i això, n'he extret una part que no empraré al motor, atés que la vull reservar com a referència per a la posterior avaluació de la traducció automàtica (*gold-standard*), com s'explica a l'apartat 3.2.3. A continuació, s'expsarà el procés seguit per al tractament del fitxer tmx de l'ECDC i del fitxer xml de la Viquipèdia.

4.2.1. EMEA

Com s'acaba d'indicar, s'han reservat 2500 línies de cada un dels fitxers per tal de no emprar-les per a l'entrenament del motor, com recomana Felipe Sánchez Martínez a l'assignatura de TAE del Màster de Tradumàtica: «entre 2.000 y 3.000 oraciones se reservan para *tuning*» (Sánchez-Martínez 2016, 55). La raó és que, a l'etapa de l'avaluació i per al *tuning* hi ha avaluadors automàtics que agafen com a referència traduccions humanes reals amb les quals no s'haja alimentat el motor. En aquest treball, no es durà a terme una etapa de *tuning*, però convé reservar part del corpus per a avaluar la qualitat dels motors o per a necessitats en el futur.

Així, s'han seleccionat a tots dos fixters el contingut de la línia 961 fins a la 3461 i s'han desat en dos fitxers per separat («EMEA_vpena.es-ro_GOLD.es» i «EMEA_vpena.es-ro_GOLD.ro»). El

corpus sense aquestes mil línies s'han desat com a «EMEA_vpena.es-ro_forMT.es» i «EMEA_vpena.es-ro_forMT.ro».

4.2.2. ECDC

Extracció de parells d'idiomes amb l'extractor Java d'ECDC

Com s'ha indicat a l'apartat 4.1.2, la memòria de traducció d'ECDC té el format tmx i inclou totes les llengües oficials de la UE. En el meu cas, interessa extraure'n el castellà i el romanés. Per a fer-ho, dins de la mateixa carpeta zip descarregada s'inclou un programa en Java («CreateLanguagePair.jar») que permet l'extracció d'un parell de llengües per a obtenir-ne una memòria de traducció tmx que incloga només aquestes dues llengües. A més a més, hi ha un fitxer d'ajuda del programa Java («Readme_CreateLanguagePair.txt»), en què s'explica com dur a terme l'extracció. Com a requisit previ, cal tenir instal·lat un client de Java per tal de poder executar l'aplicació.

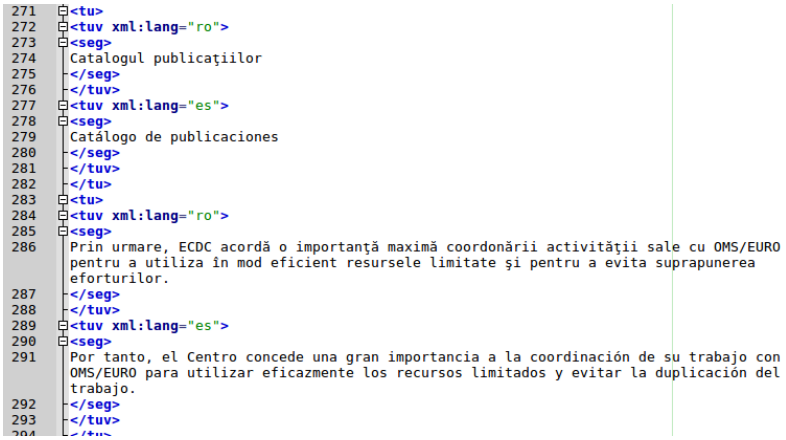
En primer lloc, cal obrir un terminal i accedir a la carpeta ECDC, on haurà de ser tant el programa Java («CreateLanguagePair.jar») com la memòria de traducció. L'ordre que cal escriure és la següent:

```
java -jar CreateLanguagePair.jar -fl First_Language -i
Input_File -o Output_File -sl Second_Language
```

En el nostre cas, l'ordre quedaria així:

```
java -jar CreateLanguagePair.jar -fl ro -i ECDC.tmx -o
ECDC_roes.tmx -sl es
```

En executar-la, s'observa que s'haurà creat un nou fitxer «ECDC_roes.tmx» a la mateixa carpeta ECDC. En obrir-la amb un editor de text, s'observa que el fitxer tmx només conté unitats de traducció en romanés i en castellà.



Il·lustració 6: fitxer tmx d'ECDC amb el parell de llengües extret.

Conversió del tmx a format Moses amb Okapi Tikal

L'Okapi Tikal[13] és eina basada en ordres multiplataforma que funciona en local. La desenvolupen ENLASO, grups d'usuaris d'Okapi Tools, Brooks Kline i l'equip del projecte Okapi. Ve inclosa dins el conjunt d'eines d'Okapi i la darrera versió (0.31) és de l'octubre del 2016. Entre les diverses opcions que ofereix, és capaç de realitzar conversions de format, com ara de tmx a Moses.

A continuació, s'explicarà com dur a terme aquesta conversió amb aquesta eina, també basada en Java. Cal dir que el mateix conjunt d'eines d'Okapi inclou una altra eina, Okapi Rainbow,[14] la qual també duu a terme conversions de fitxers, però amb interfície gràfica.

Per a fer-ho amb Okapi Tikal, en primer lloc, després d'haver descarregat les eines Okapi i descomprimir-ne la carpeta, el més fàcil és copiar el fitxer tmx a la mateixa carpeta on hi ha les eines Okapi, entre les quals hi ha Tikal. Una vegada copiat el tmx a la carpeta, cal obrir un terminal i escriure-hi l'ordre per a la conversió de tmx a Moses. En aquest cas, s'ha emprat un sistema operatiu basat en GNU/Linux, tot i aquestes eines també estan disponibles per altres sistemes com Windows.

```
sh tikal.sh -xm fitxer_entrada -2 -sl llengua_origen -tl
llengua_destí
```

Els codis de llengua han de coincidir amb els codis del fitxer tmx. En el meu cas, l'ordre ha quedat així:

13 URL: http://okapiframework.org/wiki/index.php?title=Tikal

14 URL: http://okapiframework.org/wiki/index.php?title=Rainbow

```
victor@victor-K56CM ~/MEGA/TFM/ProgramariEines/okapi-apps_gtk2-linux-x86_64_0.32
 $ sh tikal.sh -xm ECDC_roes.tmx -2 -sl ro -tl es
--------------------------------------------------------------------------------
Okapi Tikal - Localization Toolset
Version: 2.0.32
--------------------------------------------------------------------------------
Extraction to Moses InlineText
Source language: ro
Target language: es
Default input encoding: UTF-8
Filter configuration: okf_tmx
Input: /home/victor/MEGA/TFM/ProgramariEines/okapi-apps_gtk2-linux-x86_64_0.32/E
CDC_roes.tmx
Done in 1.029s
```

Il·lustració 7: conversió de tmx a moses amb Okapi Tikal.

Després de fer anar l'ordre, s'han creat dos fitxers nous: ECDC_roes.tmx.ro i ECDC_roes.tmx.es. Tanmateix, en obrir-los amb un editor de textos s'hi observa que els segments comencen i acaben amb l'etiqueta <lb/> (*line break* o salt de línia), per indicar on comença i acaba cada segment i que, a més a més, alguns tenen espais innecessaris al començament de la frase de la següent manera:

```
<lb/><lb/>                    El ECDC debe ser conocido por su
calidad y su transparencia, por los servicios que presta y
por su asesoramiento independiente.<lb/>

<lb/>Tras la exposición, el período de incubación oscila
entre dos y 30 días (con una media de 10 días).<lb/>
```

Perquè MTradumàtica processe els fitxers correctament, cal esborrar les etiquetes i els espais innecessaris. Només cal cercar i reemplaçar l'etiqueta <lb/> i els dobles espais múltiples vegades, i els fitxers quedaran nets. A més a més, també és adient cercar i substituir a tots dos fitxers el guió i l'espai que hi ha al començament d'algunes frases (-). Eines com Notepad++ permeten cercar i substituir a tots els fitxers oberts alhora, cosa molt útil, perquè es pot fer la neteja simultàniament a tots dos fitxers.

4.2.3. Viquipèdia: ús de macros a Notepad++ per al processament de fitxers xml

Un cop descarregat el fitxer xml que conté els articles seleccionats exportats, cal extraure el text adient per a entrenar el motor, tot descartant les parts de codi que no interessen per a l'entrenament. L'objectiu final és, doncs, obtenir un fitxer de text que continga una frase per línia, sense símbols ni caràcters especials, adreces web, etc. que puguen interferir en l'entrenament del motor. Finalment, el més convenient per mi va ser dur terme aquesta neteja dels fitxers XML per mitjà

de macros amb l'editor de textos lliure per a Windows Notepad++[15]. A continuació, s'explicarà com.

El primer pas va ser obrir el fitxer xml per observar-ne l'estructura. El fitxer comença amb un encapçalament sobre el tipus de fitxer (dins l'etiqueta <siteinfo>) i, a continuació, inclou cada pàgina dins l'etiqueta <page>. En observar què s'inclou dins l'etiqueta, es veu que el text rellevant és inclòs a les etiquetes <title> (el títol de l'article) i <text> (el cos de l'article).

```
<title>Abces pulmonar</title>
<ns>0</ns>
<id>787519</id>
<revision>
  <id>8262566</id>
  <parentid>8262558</parentid>
  <timestamp>2013-12-29T20:12:45Z</timestamp>
  <contributor>
    <username>Sebastianpin</username>
    <id>199625</id>
  </contributor>
  <comment>legături interne, diacritice, referințe</comment>
  <model>wikitext</model>
  <format>text/x-wiki</format>
  <text xml:space="preserve" bytes="1972">
'''Abcesul pulmonar''' este o colecție la nivelul parenchimului pulmonar, netuberculoasă,
circumscrisă. Acesta este caracterizat de prezența de spații cavitare pline cu [[puroi]]. În
cazul unor acumulări purulente numeroase (mai mici de 2 cm) se vorbește de pneumonie necrozantă.
&lt;ref name=rom&gt;[http://www.romedic.ro/abcesul-pulmonar Abcesul pulmonar] RoMedic. Accesat la
4 ianuarie 2007 &lt;/ref&gt;
```

Il·lustració 8: estructura del fitxer d'exportació de la Viquipèdia.

Així, el primer pas va ser fer una macro que busqués aquestes etiquetes i esborrés tota la resta. Això s'ha pogut fer gràcies les funcions Cerca i Comença/acaba la selecció (buscant quan comença l'etiqueta i quan acaba, i esborrant la resta de codi), enregistrant una macro que netegés un sol article, desant-la i fent-la anar fins al final del document (Macro>Executa una macro diverses vegades). El resultat de la neteja va ser el següent:

```
Abces pulmonar

'''Abcesul pulmonar''' este o colecție la nivelul parenchimului pulmonar, netuberculoasă,
circumscrisă. Acesta este caracterizat de prezența de spații cavitare pline cu [[puroi]]. În
cazul unor acumulări purulente numeroase (mai mici de 2 cm) se vorbește de pneumonie necrozantă.
&lt;ref name=rom&gt;[http://www.romedic.ro/abcesul-pulmonar Abcesul pulmonar] RoMedic. Accesat la
4 ianuarie 2007 &lt;/ref&gt;

== Condiții de apariție ==
Condițiile aspirante de apariție a abcesului pulmonar sunt:
* ''aspirarea'' conținutului orofaringian în [[plămân]] - este favorizată de dereglarea
mecanismului tusei și deglutiției. Este întâlnită în alterări ale stării de conștiență (amnezie,
```

Il·lustració 9: text extret del fitxer xml de Viquipèdia després de la primera macro.

15 El programari també es pot instal·lar a Linux amb Wine. URL: https://notepad-plus-plus.org/

S'observa que hi ha símbols que corresponen a caràcters codificats, com ara `<` o `/&gt`. Un complement de Notepad++, anomenat HTML Tag, permet descodificar aquests símbols. Per a fer-ho, cal seleccionar tot el text i anar a Complemets > HTML Tag > Decode Entities (Ctrl+Shift+E). Després del processament, hi apareixeran totes les etiquetes HTML que abans no es podien reconéixer clarament, com també altres símbols que estaven codificats.

Una vegada fet això, s'ha de fer anar una darrera macro que esborre tot això que no ens interesse i mostre el text amb una frase per línia. Cal fer anar aquesta macro tan sols una vegada i es basa exclusivament en l'ús de la funció Reemplaça combinada amb expressions regulars. A continuació, hi ha una llista en què s'explica cada pas de la macro i l'expressió de cerca utilitzada:

N.	Expressió de cerca	Reemplaça	Tipus de cerca[16]	Funció	Exemple
1	`\[[^\]\[]*|`		Expressió regular	Esborra referència d'enllaços interns de Viquipèdia.	`[[Imperiul Otoman|Imperiului Otoman]] [[Imperiului Otoman]]`
2	`<[^>]*>`		Expressió regular	Esborra totes les etiquetes HTML i XML.	`infarct pulmonar.<ref name=rom/> infarct pulmonar.`
3	`{{`	`{`	Normal	Reemplaça les claus per claus simples.	`{{reflist}}`
4	`}}`	`}`			`{reflist}`
5	`\{[^}]*\}`		Expressió regular	Esborra codi entre claus.	`{reflist|2}`
6	`|[^\n]*\n`		Expressió regular	Esborra parts de codi que no s'han pogut esborrar al pas anterior.	`| Name = Abces`
7	`}`		Normal	Esborra claus que han quedat.	
8	`https?:\/\/(www\.)?[-a-zA-Z0-9@:%._\+~#=]{2,256}\.[a-z]{2,6}\b([-a-zA-Z0-9@:%_\+.~#?&//=]*)`		Expressió regular	Esborra qualsevol URL	`http://dictionnaire.academie-medecine.fr/?q=abciximab+`
9	`[` `]`		Normal	Esborra els claudàtors	`[[puroi]] puroi`

16 Fa referència als tipus de cerca disponibles a Notepad++:
- *Normal* fa cerques que es basen exclusivament en els caràcters alfanumèrics.
- *Estés* fa cerques alfanumèriques, incloent-hi expressions per a salts de línia, de paràgraf, etc. (\n, \r, \t, \0, etc.).
- *Expressions regulars* fa cerques, com ho indica l'opció, amb l'ajuda de les expressions regulars més comunes.

10	= ’’’ ’’ :* \n* –> :-	 \n	Estés	Esborra símbols innecessaris.	== Condiţii de apariţie == Condiţiile aspirante de apariţie a abcesului pulmonar sunt: * ’’aspirarea’’ conţinutului orofaringian Condiţii de apariţie Condiţiile aspirante de apariţie a abcesului pulmonar sunt: aspirarea conţinutului orofaringian
11	.	. \n	Estés	Reemplaça l’espai que hi ha darrere d’un punt per un salt de línia (aconseguim una frase per línia).	Boala manifestă este rară şi apare la pisicile cu diverse tulburări metabolice. Se poate transmite de la o pisică la alta. Boala manifestă este rară şi apare la pisicile cu diverse tulburări metabolice. Se poate transmite de la o pisică la alta.
12	image:[^\n]* jpg[^\n]*		Expressió regular	Esborra qualsevol línia que comence per «image:» o «jpg» (restes de codi)	Image:Dornwarzen.jpg
13	Legături externe Categorie: Vezi şi		Normal	Esborra el nom de parts de l’article innecessàries que es repeteixen.	Categorie:Boli transmise sexual Boli transmise sexual
14	*		Normal	Esborra astersiscs seguits d’espai (innecessaris, apareixen a començament de línia=	* Viermii adulţi de oxiuri Viermii adulţi de oxiuri
15	\n.\n	\n	Estés	Esborra línies de punts independents.	
16	\n: \n, \n# \t	\n \n \n	Estés	Esborra línies que comencen per «: », «, » , «# ». Esborra tabuladors.	: durerile abdominale durerile abdominale

17	\n\d[^\n]*		Expressió regular	Esborra frases que comencen per xifres (normalment innecessàries)	1.
18	\n	\n	Estés		
19	\d*px		Expressió regular	Esborra parts de codi innecessàries	300px
20	ex.\n dr. \n	ex. dr.	Estés	Torna a unir frases segmentades erròniament.	Dr. Diana Florescu Dr. Diana Florescu
21	\n[A-Z]\.\n \n\.\n		Expressió regular	Esborra línies que contenen una lletra majúscula i una línia o un punt.	D. .
22	\n\n	\n	Estés	Esborra salts de línia i espais dobles (cal fer les operacions unes vuit vegades).	Doza ţintă a fost de Doza ţintă a fost de

Per al castellà, s'ha emprat la mateixa macro amb una diferència: els termes del pas 13 han estat substituïts pels equivalents en castellà («Enlaces externos», «Categoría:», «Véase también»). Una vegada tenim s'ha netejat, cal desar-lo amb format de text pla, amb l'extensió «.ro» o «.es», segons corresponga.

L'avantatge d'emprar macros a Notepad++ és que són fàcils de gravar, no cal tenir coneixements de programació i es poden exportar d'un ordinador a altre gràcies al fitxer «shortcuts.xml» que es troba als fitxers de configuració del programari.

4.3. Entrenament del motor amb MTradumàtica

En aquest apartat s'exposarà com ha estat l'experiència a la plataforma MTradumàtica[17] per a l'entrenament dels motors. S'hi explicaran els passos seguits per a entrenar-los, des de la pujada de fitxers, fins a l'obtenció del motor final. Es pot concebre com una guia d'ús d'aquesta plataforma de TAE.

4.3.1. Pujada de fitxers

17 URL: http://m.tradumatica.net

En primer lloc, cal pujar els fitxers que hem preparat a l'etapa anterior. Per a fer-ho, només cal fer clic a «Upload files» des de la pantalla d'inici. Una vegada a dins, hi ha el gestor de fitxers de la plataforma. Ara només cal arrossegar els fitxers a la part inferior de la pàgina. Després d'haver-los pujat, apareixeran a la llista de recursos, juntament amb la llengua, el nombre de línies, paraules, caràcters i la data de pujada. MTradumàtica reconeix automàticament la llengua del fitxer, tot i que en alguns casos pot fallar. Si això ocorre, cal clicar damunt de la sigla corresponent i modificar-la. Aleshores s'obri un llistat en què les llengües més probables apareixen en negreta i marcades amb un asterisc.

☐	File name	Language	Lines	Words	Chars	Date	
☐	Wikipedia-20170501094551.ro	ro	15403	213068	1397710	1/5/2017 19:00:25	
☐	EMEA_vpena.es-ro_forMT.es	es	996273	12666219	80423914	1/5/2017 19:00:24	
☐	EMEA_vpena.es-ro_forMT.ro	ro	996273	11928413	75391634	1/5/2017 19:00:23	
☐	ES-RO_ECDC.es	es	2267	40067	265173	30/4/2017 00:13:57	
☐	ES-RO_ECDC.ro	ro	2267	36472	248067	30/4/2017 00:13:57	

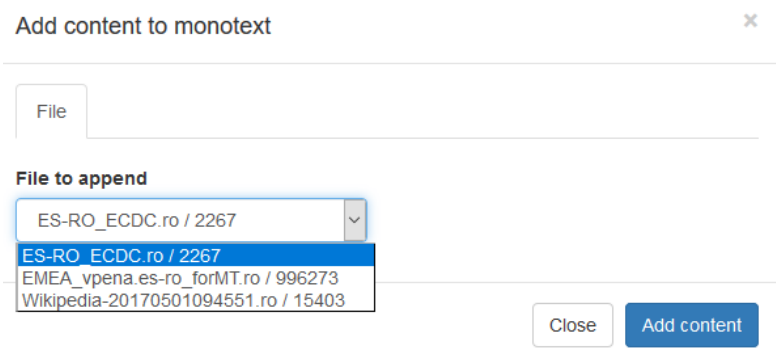
Il·lustració 10: el gestor de fitxers d'MTradumàtica amb els fitxers pujats.

Gràcies a l'última columna, es pot veure un extracte del contingut dels fitxers o descarregar-los.

4.3.2. Creació de textos monolingües

Després d'haver pujat els fitxers, cal tornar a la pantalla d'inici i fer clic a «Create Monotexts». Així, es poden utilitzar els fitxers que acabem de pujar per a crear un corpus d'entrenament del model de llengua d'arribada d'un dels motors. Una vegada a dins, cal fer clic al símbol «+» que apareixen a la primera filera de la darrera columna. S'hi mostrarà una finestra que demanarà un nom per al corpus i la llengua. En el meu cas, el nom triat ha estat «medicinafarmacia_vpi_ro», per tal de crear un corpus monolingüe en romanés.

En fer clic a «Create» s'observarà com apareix el corpus a la llista. Per a afegir-hi fitxers de text, cal fer clic al símbol «+» del corpus. Caldrà fer aquesta per a cada fitxer que vulguem que aparega al corpus monolingüe, és a dir, per a entrenar el model de llengua.

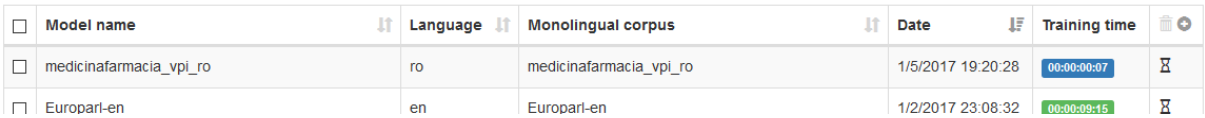


Il·lustració 11: finestra per a l'addició de fitxers de text a un corpus monolingüe.

Per conseqüent, el nombre de línies del corpus al llistat s'haurà actualitzat.

4.3.3. Elaboració de models de llengua

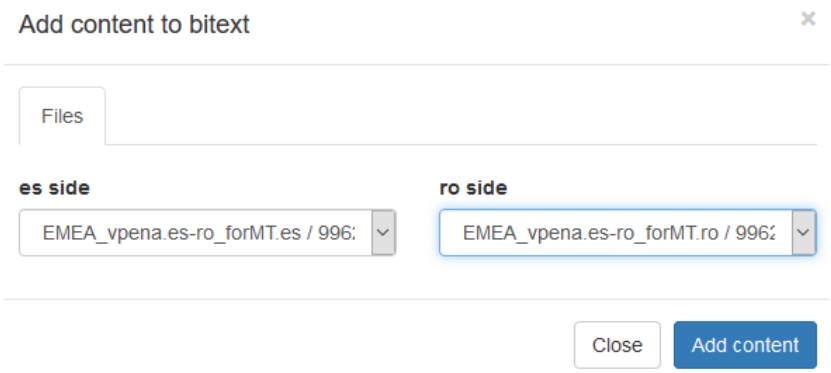
Una vegada el corpus monolingüe és a punt, ja es pot entrenar el model de llengua. Per a fer-ho, des de la pàgina d'inici farem clic a «Train Language Models». Una vegada a dins, cal fer clic al símbol «+» que apareix a la primera filera de la darrera columna. S'hi mostrarà una finestra en què cal donar nom al model de traducció, seleccionar la llengua i el corpus monolingüe de partida. En el meu cas, té el mateix nom que al corpus, «medicinafarmacia_vpi_ro». S'observarà que el motor ha començat a entrenar el model de llengua, atés que el cronòmetre de la dreta ha començat a comptar. El model de llengua estarà a punt quan passe a estar de color verd. Cal fer aquesta operació tant per al romanés com per al castellà, atés que la intenció és fer un motor en cada direcció.



Il·lustració 12: models de llengua a MTradumàtica; en blau hi ha el model en procés d'entrenament; en verd, un model entrenat.

4.3.4. Creació de textos bilingües

Una vegada s'han deixat els models de llengua d'arribada en entrenament, es pot començar a entrenar el model de traducció. En primer lloc, cal crear un corpus bilingüe a partir dels fitxers pujats. Des de la pàgina d'inici, cal fer clic a «Create a Bitext». Una vegada a dins, cal fer clic al símbol «+» que hi ha a la primera filera de la darrera columna. Hi apareixerà una finestra que demanarà un nom per al corpus i la combinació de llengües. Després de crear-lo, igual que per a generar el corpus bilingüe, caldrà seleccionar els parells de fitxers per afegir al corpus.

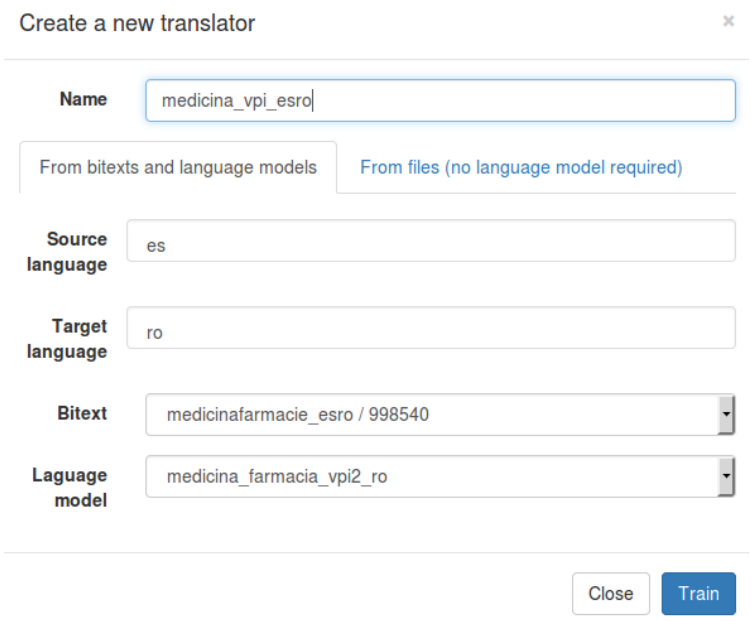


Il·lustració 13: finestra per a l'addició de fitxers de text a un corpus bilingüe.

El nombre de línies del corpus al llistat s'haurà actualitzat. Es pot fer servir un sol corpus bilingüe en totes dues direccions, és a dir, per entrenar tant el motor castellà-romanés com el motor romanés-castellà.

4.3.5. Entrenament del traductor automàtic

Una vegada s'han completat totes les fases prèvies a l'entrenament mateix del motor (model de llengua i un corpus bilingüe), ja es pot construir i entrenar el motor de TAE. Per a fer-ho, només cal anar, des de la pantalla d'inici, a «Train a Translator». Una vegada a dins, cal tornar a fer clic al símbol «+» a dalt a la dreta, i hi apareixerà de nou una finestra que demanarà la combinació de llengües, el corpus bilingüe (*bitext*) i el model de llengua.



Il·lustració 14: diàleg de creació d'un nou traductor a MTradumàtica.

Aleshores, el motor comença l'entrenament i hi apareix un cronòmetre que indica el temps que hi empra. Una vegada s'ha entrenat el motor, el cronòmetre es para i queda de color verd.

Si es vol, es pot dur a terme un ajustament automàtic dels paràmetres del sistema (*tuning*) per optimitzar el motor, com s'explica a l'apartat 3.2.1.2. Aquest procés pot ser molt lent i comporta una càrrega important al servidor. Per a fer-ho, cal fer clic a «Optimize».

Il·lustració 15: motors entrenats a MTradumàtica. S'observa que un d'ells encara es troba en la fase d'optimització.

4.4. Avaluació automàtica de la qualitat de la TA

Com s'ha explicat a l'apartat 3.2.3., és necessari avaluar la qualitat de la traducció automàtica tant a la fase d'ajustament de paràmetres del sistema, com per a l'avaluació mateixa dels resultats de traducció del motor. Aquests indicadors són molt útils per a conéixer la qualitat dels resultats, també en comparació amb altres motors.

Una eina molt útil per a dur a terme l'avaluació automàtica de la qualitat de la TA amb diversos mètodes és la plataforma Asiya. Es tracta d'un conjunt d'eines lliures per a l'avaluació automàtica de la qualitat de la TA: «an open toolkit aimed at covering the evaluation needs of system and metric developers along the development cycle» (Giménez i Màrquez 2010). Aquesta plataforma ha estat desenvolupada a la Universitat Politècnica de Catalunya, amb Lluís Màrquez, Maria Fuentes, Meritxell Gonzàlez al capdavant, i s'hi pot accedir de manera lliure per internet.[18] Permet emprar més de quinze sistemes principals d'avaluació, incloent-hi subsistemes o millores posteriors.

Preparació de fitxers

En aquest projecte, no tan sols es proposa avaluar la qualitat del motor de traducció a partir del *gold-standard*, sinó que també es pretén fer-ho a partir de textos del mateix domini d'altres procedències. Així, per a avaluar de manera automàtica la qualitat de la traducció amb Asiya s'utilitzen

- Per a l'avaluació del text extret del *gold-standard*:

18 URL: http://asiya.lsi.upc.edu/demo/asiya_online.php

- Text en la llengua de partida: s'extrau del *gold-standard* reservat del corpus d'EMEA. En aquest cas, s'han extret les 100 primeres línies tant del corpus en castellà com del corpus en romanés. Es desen en txt. Es troben a l'Annex 2 i l'Annex 3.
- Traducció humana de referència dels textos de partida original: també s'extrauen les traduccions equivalents del *gold-standard* i es desen en txt. Es troben a l'Annex 2 i l'Annex 3.
- Una o més traduccions automàtiques per valorar desades en format txt:
  - Mtradumàtica.
  - Google.
  - Yandex.
  - Apertium (només en la combinació ro>es).

- Per a l'avaluació d'un text d'una altra procedència:
  - Text en la llengua de partida:
    - Castellà (Annex 4): text extret d'un prospecte de paracetamol del web del Ministeri de Sanitat espanyol per a l'avaluació automàtica de la traducció.[19]
    - Romanés (Annex 6): Text extret d'un prospecte d'omeprazol del web Ce se întâmplă doctore per a l'avaluació automàtica de la traducció.[20]
  - Traducció humana de referència dels textos de partida original: s'han posteditat les traduccions del motor entrenat a MTradumàtica (Annex 5 i Annex 7).
  - Una o més traduccions automàtiques per valorar desades en format txt:
    - MTradumàtica.
    - Google.

    Yandex.
    - Apertium (només en la combinació ro>es).

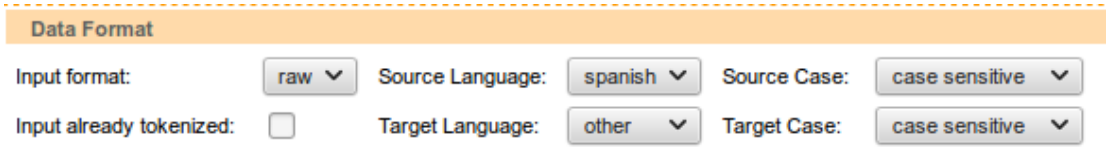
Una vegada s'han preparat els fitxers, ja es poden introduir al sistema Asiya.

Avaluació amb el sistema Asiya

19 URL: https://www.aemps.gob.es/cima/dochtml/p/69429/Prospecto_69429.html

20 URL: http://www.csid.ro/medicamente/omeprazol-terapia-20-mg-capsule-gastrorezistente-11474561/

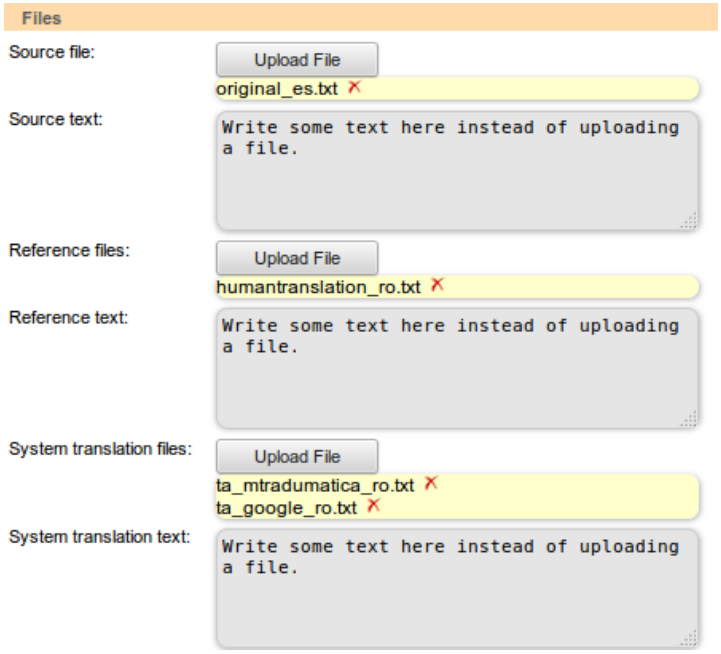
El funcionament de la plataforma Asiya és senzill i bastant intuïtiu. Cal seleccionar els paràmetres sobre les dades que introduïm: el format (xml i text pla/*raw*), la llengua de destí i d'arribada (en aquest cas *spanish* per al castelllà i *other* per al romanés), i indicar que faça diferència entre majúscules i minúscules.



Il·lustració 16: configuració del format de les dades a Asiya.

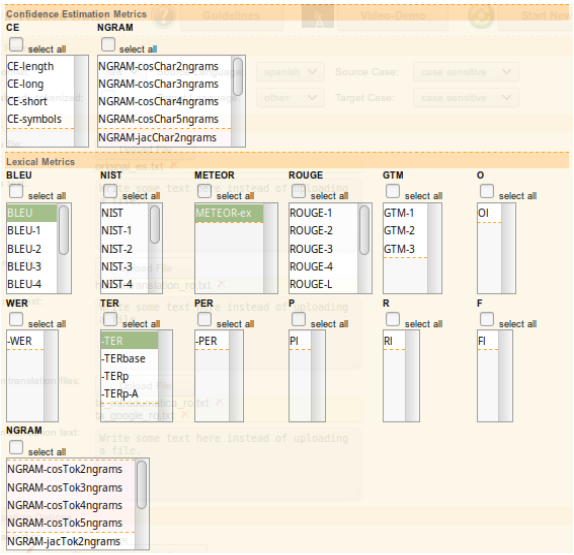
A continuació, cal seleccionar cada un dels fitxers per tal de pujar-los al sistema:

- Source file: original.
- Reference files: traducció o traduccions humanes.
- System translation files: traduccions automàtiques de cadascun dels sistemes per analitzar.



Il·lustració 17: exemple de selecció de fitxers per analitzar a Asiya.

Seguidament, se seleccionen els sistemes d'anàlisi de la qualitat de la TA fent clic a «Change metric selection»i hi apareix una pantalla en què podem seleccionar-les. En el meu cas, s'han seleccionat BLEU, METEOR-ex —*Meteor Exact Match*: una versió actualitzada de Meteor inclosa a Asiya que té en compte coincidències parcials—. i -TER —en negatiu, perquè mostra valors per davall de 0).



Il·lustració 18: selecció de sistemes d'avualuació automàtica de la TA a Asiya

Una vegada seleccionades, es fa clic a «Run Asiya!» i s'obtenen els resultats. Per defecte, es mostra l'anàlisi per segment, encara que es pot seleccionar l'anàlisi per al document i per al sistema en general (*segment level, system level i document level*). La funció *View Plot* també és interessant, atés que genera gràfics en què compara els sistemes utilitzats amb els fitxers.

4.5. Detecció i resolució d'errors

4.5.1. Error de guionets i pronoms

Després de l'anàlisi dels resultats i prèviament a la presentació dels resultats finals, es va detectar, tant fent proves aleatòries de traduccions als motors com amb paràmetres d'avaluació automàtica de la qualitat, que el traductor romanés-castellà tenia una qualitat molt inferior en comparació amb la del traductor romanés-castellà. El comportament del motor era, en molts casos, incomprensible tenint en compte la quantitat de dades amb què s'ha entrenat tant el model de llengua com el de traducció. S'observaven problemes pel que fa a la segmentació de les oracions, hi havia paraules amb pronoms que haurien de ser comunes que no traduïa, fins i tot, s'hi afegien grups de paraules inexistents a l'origen o que es repetien a cada oració, fos quina fos el text de partida. Per exemple, per al text:

«Vă rugăm să-l întrebați înainte de a utiliza acest medicament»

La traducció proposada era:

> «Si tiene cualquier otra duda sobre el să-l antes de tomar esta especialidad farmacéutica»

En observar que no traduïa la partícula «să-l», es van tornar a revisar els corpus utilitzats per a l'entrenament, amb la sorpresa que al corpus d'EMEA, el més gran, emprat tant per al model de llengua com per al model de traducció, tots els pronoms precedits d'un guionet portaven un espai al davant (să- l), de manera que el motor havia estat entrenat a partir d'un material de baixa qualitat i no reconeixia cap combinació de pronoms amb guionet.

Per tal de solucionar el problema, ha sigut necessari corregir tot el corpus. Per a fer-ho s'ha emprat la funció de Cerca i Reemplaça d'un editor de textos, en aquest cas Notepad++, per tal de reemplaçar «- » per «-». A continuació, ha calgut tornar a entrenar tots dos motors de TAE —en totes dues direccions—, seguint les passes de l'apartat 4.3. Una vegada fet això, el resultat de la traducció resulta molt millor. Així, per al mateix text de partida, la traducció que proposa és:

> «Consulte con él antes de utilizar este medicamento».

4.5.2. Error de codificació de caràcters

Com s'ha vist a l'apartat 3.4.2, l'escriptura digital del romanés no sempre s'ha fet amb la mateixa codificació de caràcters, fet que provoca que hi haja diferents codificacions per als diacrítics ș i ț, fins i tot, dins del mateix corpus. Així, com que als corpus hi podien aparéixer diferents caràcters per a la mateixa lletra, s'han entrenat dos nous motors (un en cada direcció) amb la codificació correcta d'aquests caràcters. Per a fer-ho ha calgut cercar i reemplaçar els caràcters turcs (ş i ţ) pels caràcters romanesos (ș i ț), tant en majúscules com en minúscules, a tots tres corpus en romanés:

- EMEA: tots els diacrítics eren incorrectes (turcs).
- ECDC: la majoria de diacrítics eren incorrectes (turcs).
- Viquipèdia: la majoria de diacrítics eren correctes, però en quedaven d'incorrectes.

Així, finalment s'han entrenat dos parells de motors: el primer, amb els corpus sense modificar i el segon, amb els diacrítics igualats. Es podrà veure el comportament dels motors d'MTradumàtica en cada cas.

5. Resultats

En aquest apartat s'analitzen els resultats de l'entrenament del motor en relació amb diversos aspectes que hi estan relacionats i que s'han plantejat a l'apartat 2 d'hipòtesis i objectius: l'avaluació automàtica de la qualitat, la comparació d'aquests resultats amb altres traductors automàtics, el funcionament mateix de la plataforma MTradumàtica i el comportament davant de les particularitats del romanés exposades anteriorment. Convé dir que finalment s'han entrenat quatre motors, dos en cada combinació d'idiomes, amb un parell entrenat amb la codificació antiga dels caràcters, amb un altre d'entrenat amb la versió correcta dels caràcters, com s'explica a l'apartat 4.5.2 i es veurà a l'apartat 5.3.2.

5.1. Avaluació de la qualitat

Com s'ha explicat al punt 4.4, s'ha utilitzat la plataforma Asiya per tal d'avaluar automàticament la qualitat del traductor automàtic entrenat a MTradumàtica i s'ha comparat amb altres tres traductors: Google, Yandex i Apertium (només ro>es) amb les mètriques BLEU, METEOR-ex i TER. A continuació, s'exposen els resultats a partir de la mostra extreta del *gold-standard*, a l'Annex 2 i a l'Annex 3, i dels dos prospectes descarregats d'Internet, que es troben a l'Annex 4 i a l'Annex 6. En primer lloc, s'analitza el motor romanés-castellà i, a continuació, el motor castellà-romanés.

5.1.1. Motor romanés-castellà

En observar el comportament del motor romanés-castellà d'MTradumàtica a partir de les mètriques automàtiques BLEU, METEOR-ex i -TER, els resultats de la qualitat de la traducció del *gold-standard* i el text posteditat són els següents, respectivament:

- BLEU: 0,6043/0,5438
- METEOR-ex: 0,7314/0,6934
- -TER: -0,3492/0,3199

Per tal de comparar-ne els resultats, s'han utilitzat les mateixes mètriques amb altres traductors. En el cas del traductor d'Apertium, els resultats són els següents:

- BLEU: 0,1854/0,3258
- METEOR-ex: 0,3524/0,514
- -TER: -0,6845/-0,4467

Per al traductor de Google, aquests són els resultats:

- BLEU: 0,4328/0,4045
- METEOR-ex: 0,5834/0,5884
- -TER: -0,4746/-0,3516

Per acabar, per al traductor de Yandex, els resultats són els següents:

- BLEU: 0,3518/0,5222
- METEOR-ex: 0,5433/0,665
- -TER: -0,5365/-0,2911

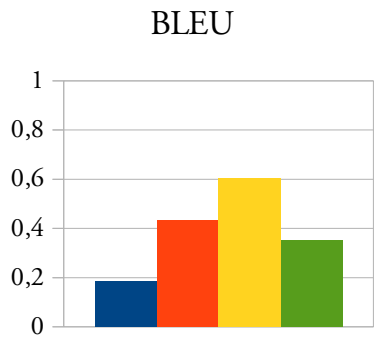
Per tant, podem observar que el motor romanés-castellà entrenat a MTradumàtica registra uns resultats prou bons que demostren la viabilitat del motor per a la postedició. A més, obté els millors resultats en relació amb la resta de motors de TA analitzats, tant en l'anàlisi de la TA del *gold-standard* com del text preeditat. La diferència amb el de Google en tots dos casos és remarcable (pel que fa a BLEU, Google se situa quasi de 1,5 a 0,2 punts per davall), encara que cal destacar que la qualitat de Yandex en el cas del text posteditat és molt semblant i que, fins i tot, registra un resultat lleugerament més bo al paràmetre -TER. D'altra banda, el resultat del motor d'Apertium és molt inferior en comparació amb la resta de motors. A continuació, es presenten els resultats en taules i en format gràfic:

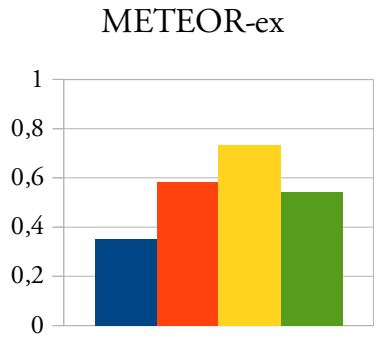
Gold-standard

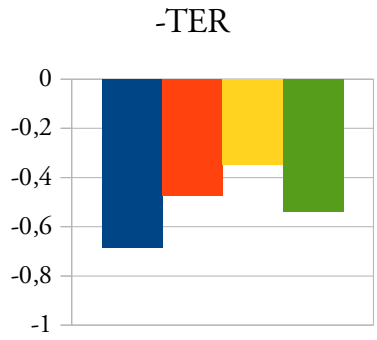
	BLEU	*METEOR-ex*	*-TER*
Apertium	0,1854	0,3524	-0,6845
Google	0,4328	0,5834	-0,4746
MTradumàtica	**0,6043**	**0,7314**	**-0,3492**
Yandex	0,3518	0,5433	-0,5365

Text posteditat

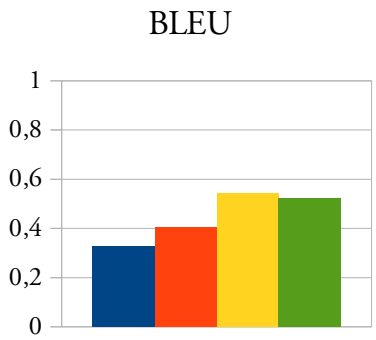
	BLEU	*METEOR-ex*	*-TER*
Apertium	0,3258	0,514	-0,4467
Google	0,4045	0,5884	-0,3516
MTradumàtica	**0,5438**	**0,6934**	-0,3199
Yandex	0,5222	0,665	**-0,2911**

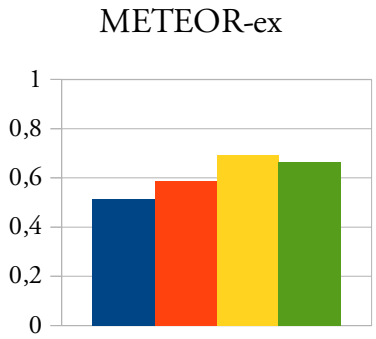


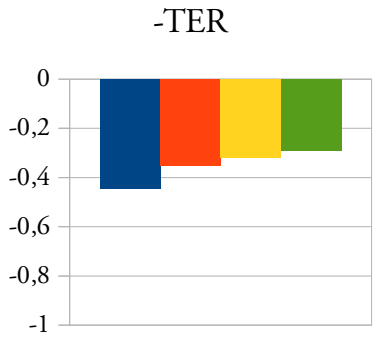




Il·lustració 19: BLEU, METEOR i -TER de l'avaluació de la TA ro>es del *gold-standard.*







Il·lustració 20: BLEU, METEOR i -TER de l'anàlisi de la TA ro>es del prospecte posteditat.

5.1.2. Motor castellà-romanés

En aquest apartat s'ha analitzat el comportament del motor castellà-romanés d'MTradumàtica amb les mètriques automàtiques BLEU, METEOR-ex i -TER. Els resultats de l'avaluació automàtica de la qualitat del *gold-standard* i el text posteditat són els següents, respectivament:

- BLEU: 0,7255/0,5399
- METEOR-ex: 0,507/0,4738
- -TER: -0,2407/-0,3436

Per tal de comparar-ne els resultats, s'han utilitzat les mateixes mètriques amb altres traductors. En el cas del traductor de Google, aquests són els resultats:

- BLEU: 0,3342/0,3463
- METEOR-ex: 0,2953/0,3044
- -TER: -0,536/-0,4359

Per acabar, per al traductor de Yandex, els resultats són els següents:

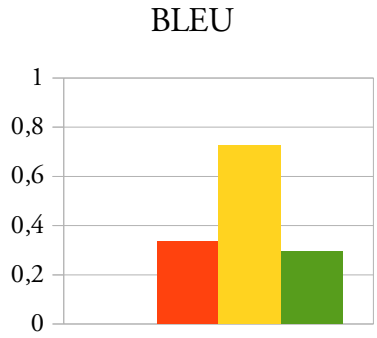
- BLEU: 0,294/0,3013
- METEOR-ex: 0,2769/0,2904
- -TER: -0,5838/-0,4923

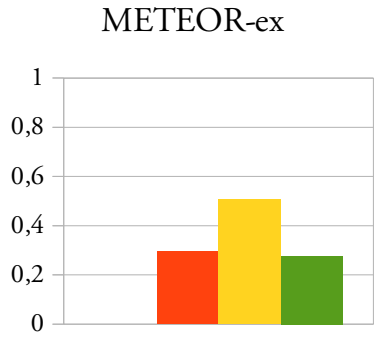
El motor castellà-romanés entrenat a MTradumàtica també registra uns resultats prou bons, semblants als del motor romanés-castellà, els quals també demostren la viabilitat del motor per a la postedició. A més, també obté els millors resultats en relació amb la resta de motors de TA analitzats, tant en l'anàlisi de la TA del *gold-standard* com del text preeditat, al marge que, en general, els resultats de Google com de Yandex són molt pobres. La diferència amb el de Google en tots dos casos és encara superior que en el cas del traductor romanés-castellà. En aquest cas, a diferència del traductor romanés-castellà, la qualitat de Yandex per al text posteditat és baixa i no gens destacable. A continuació, es presenten els resultats detallats en cadascun dels casos:

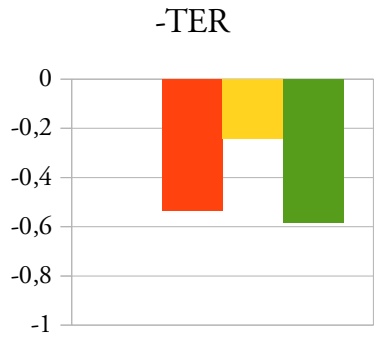
Gold standard	Text posteditat

	BLEU	*METEOR-ex*	*-TER*
Google	0,3342	0,2953	-0,536
MTradumàtica	**0,7255**	**0,507**	**-0,2407**
Yandex	0,294	0,2769	-0,5838

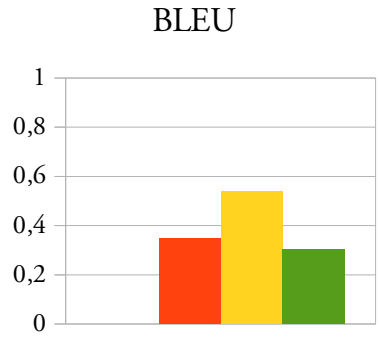
	BLEU	*METEOR-ex*	*-TER*
Google	0,3463	0,3044	-0,4359
MTradumàtica	**0,5399**	**0,4738**	**-0,3436**
Yandex	0,3013	0,2904	-0,4923

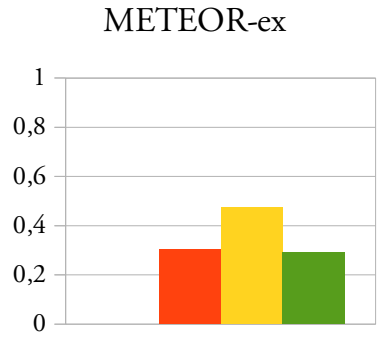


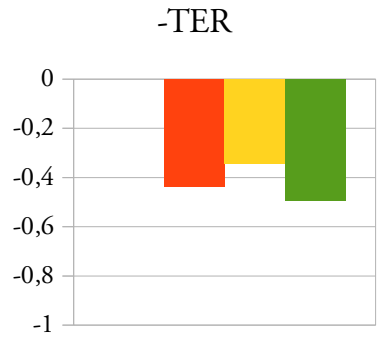




Il·lustració 21: BLEU, METEOR i -TER de l'avaluació de la TA es>ro del gold-standard.

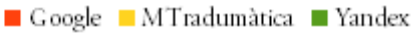






Il·lustració 22: BLEU, METEOR i -TER de l'anàlisi de la TA es>ro del prospecte posteditat.

s



5.2. Funcionament pràctic dels motors entrenats a MTradumàtica

L'objectiu d'aquest apartat és descriure el funcionament pràctic d'un motor de TAE, a partir del que s'exposa a l'apartat teòric 3.2.1. Es prendran com a referència dos exemples, un en cada llengua d'arribada, i s'hi analitzarà la divisió de l'oració en segments, les hipòtesis de traducció més probables, les traduccions més probables dels segments (a partir de la taula del model de llengua), i la versemblança en la llengua de destí (a partir del model de llengua). A més a més, al segon exemple, es veurà un cas en què hi apareix una paraula desconeguda. Per a dur a terme aquesta anàlisi s'ha utilitzat la funció Inspect[21] d'MTradumàtica

5.2.1. Exemple 1: romanés-castellà

L'oració triada per a aquest exemple és:

«Nu este recomandat tratamentul cu paracetamol în acest caz.»

S'observa als detalls de traducció que la segmentació que ha fet el motor per a l'oració és la següent:

```
[nu] = [no]

[este recomandat tratamentul] = [se recomienda el tratamiento]

[cu paracetamol] = [con paracetamol]

[în acest caz .] = [si esto ocurriera .]
```

D'altra banda, la funció Inspect, a l'apartat Translation Details, mostra les tres hipòtesis per a l'oració traduïda, juntament amb la ponderació que rep cadascun dels models estadístics que hi entren en joc, explicats l'apartat 3.2.1.1. A continuació s'exposen les tres hipòtesis de traducció més probables, amb la puntuació que el motor dóna a cada model —amb les etiquetes *LexicalReordering, Distortion, LM, WordPenalty,, etc.*— i la puntuació de la ponderació final després de combinar-los:

21 URL: http://grtradumatica.uab.cat:8080/inspect

```
0 ||| no se recomienda el tratamiento con paracetamol si esto
ocurriera . ||| LexicalReordering0= -5.73041 0 0 -0.458345 0
0 Distortion0= 0 LM0= -44.1703 WordPenalty0= -11
PhrasePenalty0= 4 TranslationModel0= -2.3235 -16.1952
-2.85713 -19.6396 ||| -2.52331

0 ||| no se recomienda el tratamiento con paracetamol con
esto . ||| LexicalReordering0= -5.73041 0 0 -0.458345 0 0
Distortion0= 0 LM0= -42.2586 WordPenalty0= -10
PhrasePenalty0= 4 TranslationModel0= -1.91804 -21.3066
-2.85713 -14.1315 ||| -2.60142

0 ||| no se recomienda el uso en este caso el tratamiento con
paracetamol . ||| LexicalReordering0= -5.59366 0 -7.59234
-0.0338495 0 -3.1381 Distortion0= -12 LM0= -41.6676
WordPenalty0= -13 PhrasePenalty0= 5 TranslationModel0=
-3.07866 -12.5144 -3.48292 -14.2567 ||| -2.6566
```

Per analitzar el funcionament del model de traducció, s'ha seleccionat un dels segments amb més d'una paraula («cu paracetamol») i s'ha cercat a l'apartat Translation Model. Així, mostra, en primer lloc, les alineacions a nivell de paraula i, a continuació, les puntuacions que reben. Es poden veure quines són les traduccions candidates per ordre de probabilitat. Aquí s'observen les tres primeres hipòtesis de traducció per a aquest segment:

```
cu paracetamol ||| paracetamol ||| 0-0 1-0 ||| 0.1875
0.110454 0.652174 0.450051

cu paracetamol ||| con paracetamol ||| 0-0 1-1 ||| 1 0.316958
0.304348 0.477073

cu paracetamol ||| con antihistamínicos y paracetamol ||| 0-0
1-1 1-3 ||| 0.25 0.159337 0.0434783 7.56746e-05
```

Per al model de traducció, la probabilitat òptima és 1. En aquest cas, s'observa que «paracetamol» és l'opció més probable, encara que la segona opció «cu paracetamol» té una mitjana de valors similars. La tercera opció és, per les xifres, la menys probable.

Pel que fa al motor de llengua, cal cercar les tres hipòtesis de traducció a l'apartat Language Model. Per a cada hipòtesi de traducció s'obté una puntuació per paraula i, a continuació, la puntua-

ció final (en aquest cas, la probabilitat òptima és 0). El paràmetre OOV mostra les paraules desconegudes (*out of vocabulary*). Aquests són els resultats per a les tres primeres hipòtesis de traducció:

```
no=487 2 -1.5218315    se=69 3 -0.419095 recomienda=7681     3
-1.011147  el=3 3 -0.78503495      tratamiento=719           3
-1.2710474 con=99 3 -0.41086942    paracetamol=9100          2
-5.527101  si=369 1 -3.8942556     esto=3107    2    -2.310963
     ocurriera=18655 3 -0.7835319  .=13 3 -1.2397515 </s>=2  3
-0.008285422     Total: -19.182915 OOV: 0

no=487 2 -1.5218315    se=69 3 -0.419095 recomienda=7681     3
-1.011147  el=3 3 -0.78503495      tratamiento=719           3
-1.2710474 con=99 3 -0.41086942    paracetamol=9100          2
-5.527101  con=99 2 -2.7475057     esto=3107    2    -3.8831625
     .=13 3 -0.7741893 </s>=2 3 -0.0017012922  Total:
-18.352684 OOV: 0

no=487 2 -1.5218315    se=69 3 -0.419095 recomienda=7681     3
-1.011147  el=3 3 -0.78503495      uso=1405   3   -0.100604564
     en=25 3 -1.7748029      este=249 3 -1.5597293   caso=527
3 -0.98204476    el=3 3 -1.9551251 tratamiento=719           3
-0.29001623      con=99 3 -0.41086942    paracetamol=9100    2
-5.527101  .=13 2 -1.7572131 </s>=2 3 -0.0014091602  Total:
-18.096025 OOV: 0
```

S'observa que les opcions que prefereixen el model de llengua i el model de traducció, per al segment seleccionat, no són la primera hipòtesi de traducció. Així, es comprova que la hipòtesi de traducció més probable és el resultat de la combinació de tots els models que formen part del motor. Això es veu a les hipòtesis de traducció de més amunt, en què es mostren els valors de cada model i la puntuació final obtinguda després de la combinació.

5.2.2. Exemple 2: castellà-romanés

L'oració triada per a aquest exemple és:

> «Enantyum es un antiinflamatorio no esteroideo potente.»

S'observa als detalls de traducció que la segmentació que ha fet el motor per a l'oració és la següent:

```
[Enantyum] = [Enantyum] UNK

[es un antiinflamatorio] = [este un medicament
antiinflamator]

[no esteroideo] = [nesteroidian]

[potente] = [puternic]

[.] = [.]
```

En aquest cas, Enantyum és una paraula desconeguda (OOV) i en respecta la e majúscula després del *truecasing*. Tot i que la paraula siga desconeguda, no sembla haver influït en la segmentació de l'oració. D'altra banda, les tres hipòtesis de traducció més probables són les següents:

```
0 ||| Enantyum este un medicament antiinflamator nesteroidian
puternic .  ||| LexicalReordering0= -0.926511 0 0 -0.759709 0
0   Distortion0=   0   LM0=   -45.0453   WordPenalty0=   -8
PhrasePenalty0=   5   TranslationModel0=   -2.62589   -5.50098
-1.04952 -6.40501 ||| -117.145

0 ||| Enantyum este un medicament antiinflamator nesteroidian
puternic .  ||| LexicalReordering0= -1.14335 0 0 -0.888459 0
0   Distortion0=   0   LM0=   -45.0453   WordPenalty0=   -8
PhrasePenalty0=   6   TranslationModel0=   -3.59773   -5.50098
-1.33648 -6.40501 ||| -117.3

0 ||| Enantyum este un medicament antiinflamator nesteroidian
puternic .  ||| LexicalReordering0= -0.952613 0 0 -0.771942 0
0   Distortion0=   0   LM0=   -45.0453   WordPenalty0=   -8
PhrasePenalty0=   6   TranslationModel0=   -3.04784   -5.50098
-2.84176 -6.405 ||| -117.399
```

S'observa que les tres traduccions coincideixen. Això passa perquè, malgrat que el motor va seguint diferents camins (hipòtesis) de traducció fins a trobar la millor, és possible que un altre camí doni el mateix resultat de traducció.

Per analitzar el funcionament del model de traducció, se selecciona un dels segments amb més d'una paraula («no esteroideo») i per veure'n quines són les traduccions candidates per ordre de probabilitat:

```
no esteroideo ||| nesteroidian ||| 0-0 1-0 ||| 0.712329
0.168397 0.945455 0.375337
```

```
no esteroideo ||| non- steroidian ||| 0-0 1-1 ||| 0.222222
0.138492 0.0363636 0.000699638

no esteroideo ||| nesteroidian " ||| 0-0 1-0 1-1 ||| 1
0.0842343 0.0181818 0.00521302
```

La probabilitat òptima és 1. En aquest cas, observem que «nesteroidian» és l'opció més probable, amb una probabilitat molt alta, atés que és la més propera a 1 amb diferència.

Pel que fa al motor de llengua, el resultat per a la hipòtesi de traducció és la següent (en aquest cas, pla probabilitat òptima és 0):

```
Enantyum=0 1 -7.617582 este=80 1 -2.2020986    un=238       2
-1.1937952 medicament=6855 3 -0.78161573 antiinflamator=12496
3 -1.4199784     nesteroidian=12497       3       -0.098959245
     puternic=5299 1 -5.1161904    .=13 2 -1.1774261 </s>=2  3
-0.002065869     Total: -19.60971 OOV: 1
```

S'observa que el model de llengua també troba una paraula desconeguda (OOV).

5.3. El cas del romanés

5.3.1. Tractament de la morfologia

Per veure quin tractament fan els motors de TAE de la morfologia romanesa, s'han agafat com a referència diverses oracions amb la paraula romanesa *antihistaminic*, per tal d'analitzar el tractament de la morfologia del romanés pel que fa a l'articulació i al fenomen dels sufixos. En primer lloc, s'han traduït amb el motor romanés-castellà diverses oracions amb aquesta paraula flexionada amb diferents articles (vegeu l'apartat 3.4.1 per a obtenir més informació sobre l'articulació del romanés).

La primera oració triada ha estat «Vitamina C este un antihistaminic natural», la traducció humana de la qual és «La vitamina C es un antihistamínico natural». En aquest cas, *antihistaminic* té l'article indefinit *un*, avantposat al nom de manera separada. El motor reconeix totes les paraules i la traducció és correcta: «La vitamina C es un antihistamínico natural».

A continuació, s'analitza què ocorre en afegir l'article definit a la paraula *antihistaminic,* la qual esdevé *antihistaminicul.* En aquest cas, la frase triada ha estat «Antihistaminicul natural este vita-

mina C», la traducció humana de la qual és «El antihistamínico natural es la vitamina C». S'observa que ara el motor no reconeix la paraula *antihistaminicul* i la deixa en romanés: «Antihistaminicul natural es la vitamina C».

Per acabar, s'ha construït l'oració amb determinant demostratiu adjectival («Vitamina C este cel mai efectiv antihistaminic natural»), la traducció humana de la qual és «La vitamina C es el antihistamínico natural más efectivo». El resultat de la TA és el següent: «La vitamina C más eficaz antihistamínico natural» En aquest cas, l'article *cel* també va avantposat, per la qual cosa reconeix la paraula *antihistaminic* i la tradueix correctament. Tot i això, no és capaç de reconstruir l'oració correctament i n'omet el verb, l'article i l'ordre de paraules.

En fer l'operació inversa, i traduir les oracions del castellà al romanés, el fenomen és diferent. Pel que fa a la primera oració («La vitamina C es un antihistamínico natural»), el motor la tradueix correctament com a «Vitamina C este un antihistaminic natural».

La segona oració («El antihistamínico natural es la vitamina C») la tradueix com a «Antihistaminice naturală este vitamina C». Com que *antihistamínico* en castellà no és una paraula desconeguda, a la traducció hi apareix traduïda, encara que amb errors: les paraules *antihistaminice* (en femení plural) i *naturală* (femení singular) hi apareixen flexionades incorrectament. Tot i això, en aquesta direcció, per la manera de flexionar del castellà, hi ha l'avantatge que hi haurà un nombre menor de paraules desconegudes, encara que les traduccions generades no tinguen la flexió correcta en romanés.

La tercera oració («La vitamina C es el antihistamínico natural más efectivo»), la tradueix com a «Vitamina C este un antihistaminic natural mai eficace». Tampoc no és capaç de respectar-ne l'estructura, encara que el resultat es podria posteditar fàcilment.

5.3.2. Problemes de codificació

Com s'ha explicat a l'apartat 3.4.2, l'estàndard per a l'escriptura dels diacrítics romanesos va arribar tard i hi ha divergències pel que fa a la codificació de la ș i la ț. Finalment, s'han entrenat dos motors: un amb els corpus amb les versions antigues dels diacrítics, i un altre amb els diacrítics actualitzats a les versions estàndard dels diacrítics.

Si fem la prova de traduir una oració amb els diacrítics correctes, observem que el motor entrenat amb les versions antigues no reconeix certes paraules. Per exemple, per a l'oració romanesa:

«Adresați-vă la doctorul dumneavoastră cât mai repede»

La traducció proposada és la següent, en què es veu que el motor no ha sabut resoldre la paraula «Adresați-vă»:

«Adresați-vă a su médico tan pronto como sea posible»

En canvi, en utilitzar el traductor amb els diacrítics actualitzats, la traducció proposada és correcta:

«Consulte a su médico lo antes posible»

Així, si es vol traduir del romanés al castellà, caldrà assegurar-se que tots els diacrítics hi apareixen en la versió correcta. Altrament, el motor no reconeixerà aquestes paraules.

Pel que fa al motor romanés-castellà, l'única diferència entre un motor i l'altre és que el traductor entrenat amb la versió incorrecta dels diacrítics produeix textos amb els diacrítics incorrectes.Conclusions

6. Conclusions

6.1. Qualitat de la TA i aplicabilitat del motor

Com s'ha vist a l'apartat 5.1., els resultats de l'avaluació automàtica de la TA en totes dues direccions demostren que la qualitat del motor és superior, i amb diferència, a la resta de motors analitzats. És destacable la qualitat del motor castellà-romanés, en què el motor entrenat amb MTradumàtica registra uns resultats BLEU superiors a Google entre 0,2 i 0,4 punts, en funció del text traduït, sempre per damunt de la qualitat mínima de 0,5 punts. S'ha de destacar que, encara que els motors amb què s'ha comparat la qualitat no estan especialitzats en el camp de la farmàcia i la medicina, és cert que disposen —en el cas dels traductors de Google i Yandex— d'una quantitat de recursos humans, tècnics i d'informació molt superiors als emprats en aquest treball.

Així, els resultats palesen diversos fets que convé destacar:

- L'especialització dels motors de TAE és una pràctica que dóna bons resultats —molt positius en aquest cas—, atés que s'aconsegueix una qualitat superior que la dels traductors privatius no especialitzats, sobretot en el cas de llengües amb menys usuaris a les quals dediquen menys recursos. Una línia de recerca futura podria ser la comparació de la qualitat dels motors amb d'altres entrenats amb corpus complementaris no pertanyents al domini.

- Es pot entrenar un motor de TAE especialitzat que done molt bons resultats amb pocs recursos. Amb una inversió mínima per a la cerca de més recursos lingüístics i la millora de la qualitat dels corpus, els resultats podrien ser superiors.

- En un cas hipotètic d'ús real, els motors serien ideals per a la traducció al sector farmacèutic i mèdic —en especial per a la traducció de prospectes—, atés que els resultats del motor són prou bons i, en conseqüència, hi podria un increment de la productivitat si es postedita la TA d'aquests motors.

6.2. Tractament de les particularitats del romanés

S'ha demostrat que particularitats del romanés exposades a l'apartat 3.4 tenen un impacte en la qualitat de la traducció automàtica estadística.

D'una banda, pel que fa a les particularitats morfològiques del romanés, es pot concloure que:

- El motor de traducció en la combinació romanés-castellà ha de fer front a paraules flexionades en romanés (amb articles sufixats, declinacions, etc.), les quals no tenen per què haver aparegut en totes les formes als materials d'entrenament. Això fa incrementar el nombre de paraules desconegudes, les quals tenen impacte en la segmentació de les oracions.
- El motor de traducció en la combinació castellà-romanés ofereix millors resultats. La raó podria ser com a hipòtesi que, per les característiques morfològiques del castellà, la flexió és més reduïda i, en conseqüència, la quantitat de paraules desconegudes és menor. Així, encara que el resultat en romanés no siga correcte pel que fa a la morfologia, el resultat és posteditable.

D'altra banda, el problema principal per a l'entrenament i l'ús de motors de TAE en romanés és la divergència pel que fa a l'ús dels diacrítics. Com es veu als apartats 3.4.2. i 4.5.2., no existeix uniformitat als textos informatitzats ni per a l'escriptura del romanés. Això fa necessari que l'usuari haja d'assegurar-se que els textos tinguen diacrítics i observar quins caràcters s'hi fan servir, tant per a l'entrenament del motor com per a la traducció, i unificar-los abans perquè el motor funcione correctament.

6.3. MTradumàtica

Després de l'experiència treballant amb MTradumàtica, s'ha pogut observar que es tracta d'una plataforma per a l'entrenament de motors de TAE completa i senzilla d'utilitzar. Així, compleix l'objectiu d'acostar la TA als traductors d'una manera intuïtiva i pedagògica: la pujada i la gestió de fitxers és molt simple, es poden crear corpus monolingües i bilingües amb pocs clics, i l'entrenament del model de llengua i dels motors és molt fàcil de configurar.

Tot i això, hi ha funcionalitats que s'hi podrien desenvolupar per tal de millorar el rendiment de la plataforma. En el cas de treballar amb el romanés, seria desitjable poder afegir-hi coneixement lingüístic, tenint en compte la morfologia i l'alt grau de derivació de la llengua. A més, pels problemes d'escriptura i representació del romanés als dispositius electrònics, seria convenient que s'hi incorporés un convertidor automàtic de caràcters que funcione automàticament abans de l'entrenament i de la traducció, juntament amb la tokenització i el *truecasing*. Així mateix, per la quantitat de textos electrònics sense diacrítics que es produeixen en romanés, fóra bo que s'hi inclogués un corrector automàtic ortogràfic i sintàctic potent, com ara LanguageTool,[22] que funcione prèviament a la traducció, que ajudés al motor a traduir textos en romanés sense diacrítics.

En definitiva, MTradumàtica és una bona plataforma i les perspectives de millora i desenvolupament són esperançadores. Les funcionalitats que es preveuen de priorització de models, compatibilitat amb altres formats, addició de coneixement lingüístic i integració amb eines TAO faran que aquesta plataforma siga accessible a traductors autònoms i empreses de traducció menudes, sense tants recursos per a dedicar a la TA.

22 URL: https://www.languagetool.org/

7. Bibliografia

Banerjee, Satanjeev, i Alon Lavie. 2005. «METEOR: An automatic metric for MT evaluation with improved correlation with human judgments». En *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 29:65–72. http://www.aclweb.org/website/old_anthology/W/W05/W05-09.pdf#page=75.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, i Andy Way. 2017. «Is Neural Machine Translation the New State of the Art?» *The Prague Bulletin of Mathematical Linguistics* 108 (1). doi:10.1515/pralin-2017-0013.

Chan, Sin-wai, ed. 2015. *The Routledge encyclopedia of translation technology*. London: Routledge.

Chunyu, Kit, i Billy Wong Tak-ming. 2015. «Evaluation in machine translation and computer-aided translation». En *The Routledge encyclopedia of translation technology*, editat per Sin-wai Chan. London: Routledge.

Daniliuc, Laura, i Radu Daniliuc. 2000. *Descriptive Romanian grammar: an outline*. München: LINCOM Europa.

Densmer, Lee. 2014. «Light and Full MT Post-Editing Explained». *Moravia Blog*. agost 19. http://info.moravia.com/blog/bid/353532/Light-and-Full-MT-Post-Editing-Explained.

Dorr, Bonnie, Joseph Olive, John McCary, i Caitlin Christianson. 2011. «Machine translation evaluation and optimization». En *Handbook of natural language processing and machine translation*, 745–843. Springer. http://link.springer.com/chapter/10.1007/978-1-4419-7713-7_5.

Forcada, Mikel. 2012. «Traducció automàtica». En *La lingüística i les seues aplicacions en la societat*, 115, 116. Publicacions de la Universitat Jaume I.

Giménez, Jesús, i Lluís Màrquez. 2010. «Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation». *The Prague Bulletin of Mathematical Linguistics* 94 (1). doi:10.2478/v10108-010-0022-6.

Ginestí, Mireia, i Mikel L. Forcada. 2009. «La traducció automàtica en la pràctica: aplicacions, dificultats i estratègies de desenvolupament». *Ginestí Rosell, Mireia ; Forcada Zubizarreta, Mikel L.. La traducció automàtica en la pràctica: aplicacions, dificultats i estratègies de desenvolupament. En: Caplletra : Revista Internacional de Filologia, 2009, No. 46: 43*. http://roderic.uv.es/handle/10550/48542.

Haddow, Barry, i Philipp Koehn. 2012. «Analysing the effect of out-of-domain data on SMT systems». En *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 422–432. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2393075.

Hearne, Mary, i Andy Way. 2011. «Statistical Machine Translation: A Guide for Linguists and Translators: SMT for Linguists and Translators». *Language and Linguistics Compass* 5 (5): 205-26. doi:10.1111/j.1749-818X.2011.00274.x.

Hutchins, Chris. 2017. «Can Neural Machine Translation Compete with Human Translation? (Part 1)». *MotionPoint*. Consulta juny 8. https://en.motionpoint.com/blog/can-neural-machine-translation-compete-with-human-translation-part-one.

KantanMT. 2017. «What is BLEU score?» *KantanMT*. Consulta maig 28. //www.kantanmt.com/.

Kaplan, Michael S. 2011. «The history of messing up Romanian on computers». *MSDN blogs*. agost 24. http://archives.miloush.net/michkap/archive/2011/08/24/10199324.html.

Koehn, Philipp. 2010. *Statistical machine translation*. Cambridge ; New York: Cambridge University Press.

Koot, David. 2016. «What's Next in MT: The N-Factor». En *Keynotes Winter 2016*, 31-33. Portland: TAUS Signature Edition. https://www.taus.net/file-downloads/download-file?path=eBooks%252FKeynotes%2BPortland%2B2016-eBook.pdf.

Martín Mor, Adrià, Ramon Piqué i Huerta, i Pilar Sánchez-Gijón. 2016. *Tradumàtica: tecnologies de la traducció*. Biblioteca de traducció i interpretació 21. Vic: Eumo.

Martín-Mor, Adrià. 2017. «MTradumàtica»: *Skase. Journal of Translation and Interpretation* 11 (1): 0025-40.

Moses. 2017. «Moses - Moses/SyntaxTutorial». Consulta febrer 26. http://www.statmt.org/moses/?n=Moses.SyntaxTutorial.

Papineni, Kishore, Salim Roukos, Todd Ward, i Wei-Jing Zhu. 2002. «BLEU: a method for automatic evaluation of machine translation». En *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1073135.

Pérez García, Isabel María. 2016. «Phrase-Based Statistical Machine Translation: Explanation of Its Processes and Statistical Models and Evaluation of the English to Spanish Translations Produced», juliol. http://rua.ua.es/dspace/handle/10045/56346.

Qun, Liu, i Zhang Xiaojun. 2015. «Machine Translation (General)». En *The Routledge encyclopedia of translation technology*, editat per Sin-wai Chan. London: Routledge.

Richens, Richard H. 1958. «Interlingual machine translation». *The Computer Journal* 1 (3): 144–147.

Sánchez-Martínez, Felipe. 2016. «Diapositives de l'assignatura Traducció Automàtica Estadística». Màster en Tradumàtica: Tecnologies de la Traducció. Universitat Autònoma de Barcelona.

Sindhu, D. V., i B. M. Sagar. 2016. «Dictionary Based Machine Translation from Kannada to Telugu». En *International Conference on Advanced Material Technologies (ICAMT)*. Vol. 2016. http://icamt2016.org/papers/MT0468.pdf.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, i John Makhoul. 2006. «A study of translation edit rate with targeted human annotation». En *Proceedings of association for machine translation in the Americas*. Vol. 200. http://mt-archive.info/AMTA-2006-Snover.pdf.

TAUS. 2017. «Knowledge Base: Pre-editing». *TAUS*. Consulta maig 22. https://www.taus.net/knowledge-base/index.php/Pre-editing.

Tiedemann, Jörg. 2012. «Parallel Data, Tools and Interfaces in OPUS». En . Instanbul. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Turovsky, Barak. 2016. «Found in translation: More accurate, fluent sentences in Google Translate». Google. *The Keyword Google Blog*. novembre 15. http://blog.google:443/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/.

———. 2017. «Higher quality neural translations for a bunch more languages». Google. *The Keyword Google Blog*. març 6. http://blog.google:443/products/translate/higher-quality-neural-translations-bunch-more-languages/.

Unicode Consortium. 2000. *The Unicode Standard, Version 3.0*. Addison-Wesley.

Yang, Liu, i Zhang Min. 2015. «Statistical Machine Translation». En *The Routledge encyclopedia of translation technology*, editat per Sin-wai Chan. London: Routledge.

Annexos

Annex 1

Instruccions pas a pas per a l'entrenament dels motors

1. Obtenció de recursos.

   (a) Descàrrega del corpus bilingüe EMEA del web del projecte OPUS[23] en format Moses. Podeu descarregar-lo des d'ací.[24]

      i. Descompressió del fitxer ZIP per a obtenir els fitxers amb extensions «.es» i «.ro».

   (b) Descàrrega del corpus multilingüe ECDC del portal EU Science Hub[25]. Podeu descarregar-lo des d'ací.[26]

      i. Descompressió del fitxer ZIP per a obtenir la memòria de traducció en format TMX multilingüe, la documentació referent a la memòria i el programari per a extraure'n parells d'idiomes.

   (c) Descàrrega d'articles de Viquipèdia.

      i. Exploració de Viquipèdia i selecció de categories i articles rellevants per al motor.

      ii. Elaboració d'un llistat de categories i articles.

      iii. Descàrrega d'articles de les categories triades en funció XML amb la funció Exporta.[27]

2. Processament dels recursos

   (a) Corpus EMEA:

23 URL: http://opus.lingfil.uu.se/

24 URL: http://opus.lingfil.uu.se/download.php?f=EMEA%2Fes-ro.txt.zip

25 URL: https://ec.europa.eu/jrc/en/language-technologies

26 URL: http://langtech.jrc.ec.europa.eu/Resources/ECDC-TM/ECDC-TM.zip

27 URL: https://ro.wikipedia.org/wiki/Special:Export%C4%83

i. Conversió de diacrítics romanesos a la codificació correcta (apartat 4.5.2). Amb un editor de textos com Notepad++, amb la funció Cerca i Reemplaça: Ş per Ș, Ţ per Ț, ş per ș i ţ per ț. Per a fer-ho amb més rapidesa i amb altres recursos, s'ha gravat una macro que realitza les quatre substitucions.[28] El nom de la macro al paquet és «4_replace_romanian_characters».

ii. Cerca i reemplaça de «- » per «-», per tal d'unir els pronoms als verbs i corregir l'error explicat a l'apartat 4.5.1.

iii. Reserva de 2.500 línies per a una *gold-standard* del corpus, per a *tuning* o avaluar-ne la qualitat. Extracció de les mateixes 2.500 línies de cada llengua amb Notepad++, desant el text extret en un fitxer per a cada llengua. Finalment, han de quedar dos fitxers per a cada llengua (un per a l'entrenament i un altre per a l'avaluació).

(b) Corpus ECDC.

i. Extracció del parell d'idiomes amb l'extractor Java inclòs a la carpeta descarregada. Es necessita tenir un client Java instal·lat per fer anar el programa amb el terminal d'ordres.

ii. Conversió del tmx bilingüe a format Moses. Es pot utilitzar Okapi Tikal (basat en el terminal de comandes) o Okapi Rainbow (amb interfície visual), de la suite Okapi.[29]

iii. Neteja d'etiquetes de separació de segments </lb> en un editor de textos. Cal cercar i reemplaçar aquesta etiqueta per tal d'esborrar-la totes les vegades que apareix al fitxer.

---

28 El paquet de macros gravades per al projecte es poden descarregar lliurement des de l'adreça web http://bit.ly/2qVbNpF. Per a instal·lar-les, només cal reemplaçar el fitxer «shortcuts.xml», el qual es troba a Windows a %AppData%\Notepad++\. Trobareu més informació a: https://stackoverflow.com/questions/5431964/where-are-the-recorded-macros-stored-in-notepad

29 URL: http://okapiframework.org/wiki/index.php?title=Main_Page

iv. Per al romanés, unificació de caràcters per als diacrítics a Notepad++ amb la macro «4_replace_romanian_characters».

(c) Viquipèdia

i. Extracció del text útil a Notepad++ amb la macro «1_keep_title_text», per tal de conservar només allò que hi ha dins les etiquetes <title> i <text>. Cal repetir la macro fins al final del document.

ii. Descodificació d'entitats HTML amb el complement HTML Tag de Notepad++.

iii. Neteja del corpus a Notepad++ amb la macro «2_final_macro_clean_ro» o «3_final_macro_clean_es», en funció de la llengua.

iv. Per al romanés, unificació de caràcters per als diacrítics a Notepad++ amb la macro «4_replace_romanian_characters».

v. Conversió dels textos netejats a format de text pla, desant els fitxers a Notepad++ amb les extensions «es» o «ro», segons corresponga.[30]

3. Entrenament del motor a MTradumàtica[31]

(a) Pujada dels fitxers de text amb les extensions «.es» i «.ro», segons corresponga, a la pàgina «File manager».[32]

(b) Creació de corpus monolingües en cada llengua (es i ro) per a l'entrenament del model de llengua, a la pàgina «Monotext manager»[33]. Cal triar un nom per al corpus i afegir-hi tots els fitxers de text per al model de llengua (en aquest cas, EMEA, ECDC i Viquipèdia).

30 Els corpus monolingües de textos extrets de Viquipèdia es poden descarregar lliurement des de l'enllaç http://bit.ly/2r2nE08.

31 URL: http://m.tradumatica.net

32 URL: http://grtradumatica.uab.cat:8080/files

33 URL: http://grtradumatica.uab.cat:8080/monolingual_corpora

(c) Entrenament dels models de llengua a la pàgina «Language model trainer».[34] Cal triar un nom i afegir el corpus monolingüe creat a l'apartat anterior. S'ha de fer per a cada llengua.

(d) Creació d'un corpus bilingüe (es<>ro) per a l'entrenament dels motors a la pàgina «Bi-text manager»[35]. Cal triar un nom per al corpus i afegir-hi tots els fitxers de text paral·lels per al model de traducció (en aquest cas, EMEA i ECDC).

(e) Entrenament dels motors a la pàgina «Translator Trainer».[36] Cal posar un nom al traductor, seleccionar el corpus bilingüe, la combinació de llengües i el model de la llengua d'arribada. Una vegada entrenat, a la mateixa pàgina es pot optimitzar el motor (*tuning*) fent clic a Optimize.

(f) Traducció amb els motors a la pàgina «Translate». Els motors ja apareixen al llistat de motors disponibles a MTradumàtica.

34 URL: http://grtradumatica.uab.cat:8080/language_models

35 URL: http://grtradumatica.uab.cat:8080/bitexts

36 URL: http://grtradumatica.uab.cat:8080/translators

Annex 2

Extracte del gold-standard *per a l'avaluació automàtica de la traducció (es)*

La mejoría clínica del paciente durante el tratamiento antipsicótico, puede tardar entre varios días a algunas semanas.
Los pacientes deben estar estrechamente controlados durante este periodo.
La aparición de comportamiento suicida es inherente a las patologías psicóticas y trastornos del estado de ánimo y en algunos casos han sido notificados algunos casos tempranos tras la administración inicial o cambio de terapias antipsicóticas, incluyendo el tratamiento con aripiprazol (ver sección 4.8).
Debe realizarse una supervisión estrecha de los pacientes de alto riesgo cuando son tratados con antipsicóticos. Los resultados de un estudio epidemiológico demostraron que no se incrementó el riesgo de suicidio con aripiprazol comparado con otros antipsicóticos entre pacientes con trastorno bipolar.
Alteraciones cardiovasculares: aripiprazol debe ser utilizado con precaución en pacientes que presentan enfermedad cardiovascular conocida (historia de infarto de miocardio o enfermedad isquémica cardiaca, fallo cardíaco, o anormalidades de la conducción), enfermedad cerebrovascular, condiciones en las que puede predisponerse a pacientes a la hipotensión (deshidratación, hipovolemia, y tratamiento con medicación antihipertensiva) o hipertensión, incluyendo acelerada o maligna.
Anormalidades de la conducción: en ensayos clínicos de aripiprazol, la incidencia de la prolongación del intervalo QT fue comparable al placebo.
Como con otros antipsicóticos, aripiprazol debe ser utilizado con precaución en pacientes con historia familiar de prolongación del intervalo QT.
Discinesia tardía: en ensayos clínicos de hasta un año de duración, se han notificado casos poco frecuentes de discinesia durante el tratamiento con aripiprazol.
Si aparecen signos y síntomas de discinesia tardía en un paciente tratado con ABILIFY, se debe considerar una reducción de dosis o la interrupción del tratamiento.
Estos síntomas pueden empeorar temporalmente o incluso pueden manifestarse después de la interrupción del tratamiento.
27 Síndrome Neuroléptico Maligno (SNM):
SNM es un complejo de síntomas potencialmente mortal asociado con fármacos antipsicóticos.
En ensayos clínicos se han notificado casos raros de SNM durante el tratamiento con aripiprazol.
Las manifestaciones clínicas del SNM son hipertermia, rigidez muscular, estado mental alterado y evidencia de inestabilidad autonómica (pulso o presión sanguínea irregular, taquicardia, diaforesis y disritmia cardíaca).
Los signos adicionales pueden incluir un aumento de creatina fosfoquinasa, mioglobinuria (rabdomiólisis), e insuficiencia renal aguda.
También se han notificado elevaciones de la creatina fosfoquinasa y rabdomiólisis no necesariamente asociadas con el SNM.
Si un paciente desarrolla signos y síntomas indicativos de SNM, o presenta una fiebre alta inexplicable sin manifestaciones clínicas adicionales de SNM, todos los medicamentos antipsicóticos, incluyendo ABILIFY deben ser interrumpidos.

Convulsiones: en ensayos clínicos se han notificado casos poco frecuentes de convulsiones durante el tratamiento con aripiprazol.
Por lo tanto, se debe utilizar aripiprazol con precaución en pacientes epilépticos o con historia de convulsiones.
Pacientes de edad avanzada con psicosis relacionada con demencia:
Aumento de mortalidad: en tres ensayos controlados con placebo (n= 938; edad media:
82,4 años; rango:
56-99 años) de aripiprazol en pacientes de edad avanzada con psicosis asociada a la enfermedad de Alzheimer, los pacientes tratados con aripiprazol presentaron mayor riesgo de muerte comparado con placebo.
El porcentaje de muerte en pacientes tratados con aripiprazol fue del 3,5% comparado con el 1,7% del grupo placebo.
Aunque las causas de muerte fueron variadas, la mayoría de las muertes parecieron ser de naturaleza cardiovascular (por ejemplo, fallo cardíaco, muerte súbita) o infecciosa (por ejemplo, pneumonía).
Reacciones adversas cerebrovasculares: en los mismos ensayos, se notificaron reacciones adversas cerebrovasculares (por ejemplo ictus, crisis isquémica transitoria), incluyendo casos con desenlace fatal (media de edad:
84 años; rango:
78-88 años).
En total, en estos ensayos, un 1,3% de los pacientes tratados con aripiprazol notificaron reacciones adversas cerebrovasculares en comparación con un 0,6% de los pacientes tratados con placebo.
Esta diferencia no fue estadísticamente significativa.
Sin embargo, en uno de estos ensayos, un ensayo de dosis fijas, hubo una relación dosis respuesta significativa para las reacciones adversas cerebrovasculares en pacientes tratados con aripiprazol.
ABILIFY no está aprobado para el tratamiento de la psicosis relacionada con demencia.
Hiperglucemia y Diabetes Mellitus: se ha notificado hiperglucemia, en algunos casos extrema y asociada con cetoacidosis o coma hiperosmolar o muerte, en pacientes tratados con antipsicóticos atípicos, incluyendo ABILIFY.
Los factores de riesgo que pueden predisponer a los pacientes a padecer complicaciones graves incluyen la obesidad y la historia familiar de diabetes.
En los ensayos clínicos con aripiprazol, no hubo diferencias significativas en la tasa de incidencia de reacciones adversas hiperglucémicas (incluyendo diabetes) en los valores clínicos anómalos de glucemia en comparación con placebo.
No se dispone de la valoración del riesgo exacto de reacciones adversas hiperglucémicas que permita la comparación directa entre pacientes tratados con ABILIFY y pacientes tratados con otros antipsicóticos atípicos.
Se debe prestar atención a los signos y síntomas de hiperglucemia (como polidipsia, poliuria, polifagia y debilidad), en pacientes tratados con antipsicóticos, incluyendo ABILIFY, y se debe monitorizar regularmente en pacientes con diabetes mellitus o con factores de riesgo de diabetes mellitus por el empeoramiento del control de glucosa.
Aumento de peso: el aumento de peso se ve comúnmente en pacientes con esquizofrenia y manía bipolar debido a la co-morbilidad, uso de antipsicóticos conocidos que causan aumento de peso, estilo de vida poco saludable, y puede conducir a complicaciones graves.

El aumento de peso ha sido notificado en la post- comercialización entre pacientes a los que se les prescribe ABILIFY.
Cuando se han visto, han sido habitualmente en aquellos con factores significativos de riesgo tales como historia de diabetes, alteraciones tiroideas o adenoma de la pituitaria.
En ensayos clínicos con aripiprazol no se ha mostrado que induzcan a aumento de peso clínicamente relevante (ver sección 5.1).
Disfagia: dismotilidad esofágica y aspiración han sido asociadas con la utilización de antipsicóticos incluyendo ABILIFY.
Aripiprazol y otros medicamentos antipsicóticos deben ser utilizados con precaución en pacientes con riesgo de neumonía por aspiración.
28 Lactosa: los pacientes con problemas hereditarios poco frecuentes de intolerancia a la galactosa, insuficiencia de lactasa o malabsorción de glucosa-galactosa no deben tomar este medicamento.
Hipersensibilidad: como con otros medicamentos pueden aparacer reacciones de hipersensibilidad con aripiprazol, caracterizadas por síntomas alérgicos (ver sección 4.8).
4.5 Interacción con otros medicamentos y otras formas de interacción
Debido al antagonismo del receptor α 1-adrenérgico, aripiprazol puede aumentar los efectos de ciertos agentes antihipertensivos.
Debido a la acción principal de aripiprazol en el SNC, se debe utilizar aripiprazol con precaución cuando se utilice en combinación con alcohol u otros medicamentos de acción central, ya que se solapan las reacciones adversas como sedación (ver sección 4.8).
Debe utilizarse con precaución la administración concomitante de aripiprazol con otros fármacos que produzcan prolongación del intervalo QT o trastornos electrolíticos.
Interacción de otros medicamentos con ABILIFY:
Famotidina, antagonista H2, bloqueante de los ácidos gástricos, reduce la tasa de absorción de aripiprazol, pero se considera que este efecto no es clínicamente significativo.
Aripiprazol se metaboliza por múltiples vías involucrando los enzimas CYP2D6 y CYP3A4 pero no el enzima CYP1A.
Por lo tanto, no es necesario un ajuste de dosis en fumadores.
En un ensayo clínico en sujetos sanos, un potente inhibidor de CYP2D6 (quinidina) aumentó el AUC de aripiprazol en 107%, mientras que la Cmax no varió.
El AUC y la Cmax de dehidro-aripiprazol, el metabolito activo, disminuyeron un 32% y un 47%, respectivamente.
La dosis de ABILIFY se debe reducir a la mitad de la dosis prescrita si se administra ABILIFY conjuntamente con quinidina.
Otros potentes inhibidores de CYP2D6, como fluoxetina o paroxetina, posiblemente tengan efectos similares y se deben administrar, por lo tanto, reducciones similares de dosis.
En un ensayo clínico en sujetos sanos, un potente inhibidor de CYP3A4 (ketoconazol) aumentó el AUC y la Cmax de aripiprazol en 63% y 37%, respectivamente.
Aumentó el AUC y la Cmax de dehidro-aripiprazol en 77% y 43%, respectivamente.
En metabolizadores lentos de CYP2D6, el uso concomitante de potentes inhibidores de CYP3A4 puede aumentar las concentraciones plasmáticas de aripiprazol comparado con metabolizadores rápidos de CYP2D6.
Cuando se considere la administración concomitante de ketoconazol u otro potente inhibidor CYP3A4 con ABILIFY el posible beneficio debe estar por encima del posible riesgo para el paciente.

Si se administra ketoconazol junto con ABILIFY, la dosis de ABILIFY se debe reducir a la mitad de la dosis prescrita.
Con otros potentes inhibidores de CYP3A4 como itraconazol e inhibidores de la proteasa VIH, se pueden esperar efectos similares y se deben administrar, por lo tanto, reducciones similares de dosis.
En caso de interrupción del inhibidor CYP2D6 ó 3A4, se debe aumentar la dosis de ABILIFY hasta el nivel anterior a la iniciación del tratamiento concomitante.
Se pueden esperar pequeños aumentos de las concentraciones de aripiprazol cuando se utiliza de forma concomitante con inhibidores débiles del CYP3A4 (por ejemplo, diltiazem o escitalopram) o del CYP2D6.
Después de una administración concomitante con carbamazepina, un potente inductor de CYP3A4, las medias geométricas de Cmax y AUC para aripiprazol fueron 68% y 73% más bajas, respectivamente, si se compara con la administración de aripiprazol (30 mg) sólo.
Del mismo modo, las medias geométricas de Cmax y AUC para dehidro-aripiprazol después de la co-administración de carbamazepina fueron 69% y 71% más bajas, respectivamente, que las de un tratamiento con aripiprazol sólo.
29 La dosis de ABILIFY debe ser duplicada cuando se administra conjuntamente con carbamazepina.
De otros potentes inductores de CYP3A4 (tales como rifampicina, rifabutina, fenitoina, fenobarbital, primidona, efavirenz, nevirapina y Hierba de San Juan (Hypericum perforatum) se deben esperar efectos similares y se deben administrar, por lo tanto, aumentos similares de dosis.
En cuanto se suspendan los inductores potentes de CYP3A4 la dosis de ABILIFY debe ser reducida a la dosis recomendada.
No se han encontrado cambios clínicamente significativos si se administra conjuntamente aripiprazol con valproato o litio.
Interacción de ABILIFY con otros medicamentos:
En ensayos clínicos, dosis de aripiprazol de 10-30 mg/ día no tuvieron efectos significativos sobre el metabolismo de los substratos CYP2D6 (índice dextrometorfano/ 3-metoximorfinano), 2C9 (warfarina), 2C19 (omeprazol) y 3A4 (dextrometorfano).
Además aripiprazol y dehidro-aripiprazol no modifican el metabolismo mediado por CYP1A2, in vitro.
Por lo tanto, es improbable que aripiprazol provoque interacciones medicamentosas clínicamente importantes mediadas por estos enzimas.
Cuando se administra aripiprazol conjuntamente con valproato o litio, no se han encontrado cambios clínicamente significativos en las concentraciones de valproato o litio.
4.6 Embarazo y lactancia
No hay ensayos en mujeres embarazadas bien controlados y adecuados con aripiprazol.
Estudios en animales, no pueden excluir el desarrollo potencial de toxicidad (ver sección 5.3).
Las pacientes deben notificar a su médico si están embarazadas o tienen intención de quedarse embarazadas durante el tratamiento con aripiprazol.
Debido a información de seguridad insuficiente en humanos y datos inciertos en estudios de reproducción animal, este medicamento no debe utilizarse en el embarazo, a menos que el beneficio esperado justifique claramente un riesgo potencial en el feto.
Aripiprazol se excreta en la leche de ratas tratadas durante la lactancia.
Se desconoce si aripiprazol se excreta en la leche humana.

Se debe advertir a las pacientes que no den el pecho si están tomando aripiprazol.
4.7 Efectos sobre la capacidad para conducir y utilizar máquinas
No se han realizado estudios de los efectos sobre la capacidad para conducir y utilizar máquinas.
Sin embargo, al igual que con otros antipsicóticos, los pacientes deben tener cuidado al utilizar máquinas peligrosas, incluyendo vehículos a motor, hasta que tengan la certeza de que aripiprazol no les afecta negativamente.
4.8 Reacciones adversas
Las siguientes reacciones adversas se producen con más frecuencia (≥ 1/ 100) que con placebo, o fueron identificadas como reacciones adversas de posible relevancia médica (*):
La frecuencia de reacciones adversas listadas a continuación se definen siguiendo la siguiente clasificación: frecuentes (≥ 1/ 100, < 1/ 10) y poco frecuentes (≥ 1/ 1,000, < 1/ 100).
Poco frecuentes: hipotensión ortostática* Trastornos generales y alteraciones en el lugar de administración:
Frecuentes: agitación, insomnio, ansiedad
Síntomas extrapiramidales (SEP):
Esquizofrenia - en un ensayo controlado a largo plazo de 52 semanas, los pacientes tratados con aripiprazol tuvieron una incidencia global menor (25,8%) de SEP incluyendo parkinsonismo, acatisia, distonía y discinesia, comparados con aquellos tratados con haloperidol (57,3%).
En un ensayo controlado con placebo a largo plazo, de 26 semanas, la incidencia de SEP fue de 19% para pacientes tratados con aripiprazol y 13,1% para pacientes tratados con placebo.
En otro ensayo controlado a largo plazo de 26 semanas, la incidencia de SEP fue de 14,8% para pacientes tratados con aripiprazol y 15,1% para pacientes tratados con olanzapina.
Episodios maníacos en Trastorno Bipolar I - en un ensayo controlado de 12 semanas de duración, la incidencia de SEP fue de 23,5% en pacientes tratados con aripiprazol y de 53,3% en pacientes tratados con haloperidol.
En otro ensayo, también de 12 semanas de duración, la incidencia de SEP fue de 26,6% en pacientes tratados con aripiprazol y de 17,6% para aquellos tratados con litio.
En la fase de mantenimiento a largo plazo de 26 semanas de duración de un ensayo controlado con placebo, la incidencia de SEP fue de 18,2% en pacientes tratados con aripiprazol y de 15,7% en pacientes tratados con placebo.
En ensayos controlados con placebo, la incidencia de acatisia en pacientes con trastorno bipolar fue de 12,1% en los tratados con aripiprazol y de 3,2% en aquellos que recibieron placebo.

Annex 3

Extracte del gold-standard *per a l'avaluació automàtica de la traducció (ro)*

În timpul tratamentului antipsihotic, îmbunătățirea stării clinice a pacientului se poate produce după câteva zile până la câteva săptămâni.
Pacienții trebuie monitorizați atent pe parcursul acestei perioade.
Apariția comportamentului suicidar este inerent în cazul afecțiunilor psihotice și a tulburărilor de dispoziție în unele cazuri s-a raportat precoce după inițierea sau schimbarea terapiei antipsihotice, inclusiv a tratamentului cu aripiprazol (vezi pct 4. 8).
Rezultatele unui studiu epidemiologic au arătat că nu există un risc crescut de suicid cu aripiprazol comparativ cu alte antipsihotice, la pacienții cu tulburare bipolară.
Tulburări cardiovasculare: la pacienții cu afecțiuni cardiovasculare (antecedente de infarct miocardic sau boală cardiacă ischemică, insuficiență cardiacă sau tulburări de conducere), afecțiuni cerebrovasculare, condiții care predispun la hipotensiune (deshidratări, hipovolemie și tratament cu medicamente antihipertensive) sau hipertensiune inclusiv forma cu evoluție accelerată sau malignă, aripiprazolul trebuie utilizat cu precauție.
Tulburări de conducere: incidența intervalului QT prelungit în studiile clinice cu aripiprazol a fost comparabilă cu placebo.
La pacienții cu istoric familial de QT prelungit, ca și în cazul altor antipsihotice aripiprazolul trebuie utilizat cu precauție.
Dischinezie tardivă: în studiile clinice cu durata de cel mult un an, au existat raportări mai puțin frecvente de dischinezie determinată de tratamentul cu aripiprazol.
Dacă la pacienții tratați cu ABILIFY apar semne și simptome de dischinezie tardivă, trebuie avută în vedere reducerea dozei sau întreruperea administrării.
Aceste simptome se pot agrava temporar sau chiar pot să apară după întreruperea tratamentului.
Sindrom neuroleptic malign (SNM):
SNM este un sindrom complex, potențial letal, asociat administrării medicamentelor antipsihotice.
În studiile clinice, în timpul tratamentului cu aripiprazol s -
Manifestările clinice ale SNM sunt hiperpirexia, rigiditatea musculară, alterarea statusului mental și semne de instabilitate vegetativă (puls neregulat sau tensiune arterială oscilantă, tahicardie, diaforeză și tulburări cardiace de ritm).
Alte semne pot include creșterea valorii creatin fosfokinazei, mioglobinurie (rabdomioliză) și insuficiență renală acută.
Cu toate acestea, s-au raportat creșteri ale creatin fosfokinazei și rabdomioliză, nu neaparat în asociere cu SNM.
Dacă un pacient dezvoltă semne și simptome caracteristice pentru SNM sau prezintă febră foarte mare, inexplicabilă, fără alte manifestări clinice de SNM, trebuie întreruptă administrarea tuturor medicamentelor antipsihotice, inclusiv a ABILIFY.
Convulsii: în studiile clinice, în timpul tratamentului cu aripiprazol s-au raportat cazuri mai puțin frecvente de convulsii.
Ca urmare, aripiprazolul trebuie utilizat cu precauție la pacienții cu antecedente de tulburări convulsive sau cu afecțiuni asociate cu convulsiile.

Pacienți vârstnici cu psihoze asociate demenței:
Mortalitate crescută: în trei studii clinice controlate cu placebo (n= 938; vârsta medie:
82, 4 ani; interval:
56-99 ani), efectuate cu aripiprazol la pacienți vârstnici cu psihoze asociate cu boala Alzheimer, pacienții tratați cu aripiprazol au prezentat risc crescut de deces, comparativ cu placebo.
La pacienții tratați cu aripiprazol, frecvența decesului a fost de 3, 5%, comparativ cu 1, 7% în grupul placebo.
Deși cauzele de deces au variat, majoritatea au fost fie de cauză cardiovasculară (de exemplu: insuficiență cardiacă, moarte subită), fie infecțioasă (de exemplu, pneumonie).
Evenimente adverse cerebrovasculare: în aceleași studii clinice, la pacienți (vârsta medie:
84 ani; interval:
78-88 ani) s-au raportat evenimentele adverse cerebrovasculare (de exemplu: accident vascular cerebral, accident ischemic tranzitor), incluzând decese.
În ansamblu, în aceste studii, la 1, 3% dintre pacienții tratați cu aripiprazol s-au raportat evenimente adverse cerebrovasculare, comparativ cu 0, 6% dintre pacienții cărora li s-a administrat placebo.
Această diferență nu a fost semnificativă statistic.
Cu toate acestea, într-unul dintre aceste studii, un studiu cu doză fixă, la pacienții tratați cu aripiprazol a existat o relație semnificativă dependentă de doză a evenimentelor adverse cerebrovasculare.
ABILIFY nu este aprobat pentru tratamentul psihozelor asociate demenței.
Hiperglicemie și diabet zaharat: la pacienții tratați cu antipsihotice atipice, inclusiv cu ABILIFY s-a raportat hiperglicemie, în unele cazuri marcată și asociată cu cetoacidoză și comă hiperosmolară sau deces.
Obezitatea și antecedentele familiale de diabet zaharat, sunt unii dintre factorii de risc ce ar putea predispune pacienții la coplicații severe.
În studiile clinice cu aripiprazol, nu au existat diferențe semnificative între frecvențele incidenței hiperglicemiei asociate evenimentelor adverse (incluzând diabetul zaharat) sau ale valorilor de laborator anormale ale glicemiei, comparativ cu placebo.
Nu este disponibil un risc precis estimat pentru hiperglicemia asociată evenimentelor adverse la pacienții tratați cu ABILIFY și alte antipsihotice atipice, pentru a permite comparații directe.
Pacienții tratați cu orice antipsihotice, incluzând ABILIFY, trebuie supravegheați pentru a se observa semnele și simptomele de hiperglicemie (cum sunt: polidipsie, poliurie, polifagie și slăbiciune), iar pacienții cu diabet zaharat sau cu factori de risc pentru diabet zaharat trebuie monitorizați regulat pentru a se observa reducerea controlului glucozei.
Creșterea în greutate: creșterea în greutate este frecvent întâlnită la pacienții cu schizofrenie și manie bipolară datorită co-morbidităților, a utilizării antipsihoticelor cunoscute a determina creșteri în greutate, a stilului de viață dezordonat, și ar putea determina complicații severe.
În perioada post -autorizare, printre pacienții tratați cu ABILIFY, creșterea în greutate a fost raportată.
Atunci când este întâlnită, apare mai ales la cei cu factori de risc semnificativi, precum antecedente de diabet, afecțiuni ale tiroidei sau adenom de glandă pituitară.
5. 1).
Disfagia: utilizarea medicamentelor antipsihotice, inclusiv a ABILIFY, s-a asociat cu afectarea motilității esofagiene si aspirație.

Aripiprazolul și alte medicamente antipsihotice, trebuie utilizate cu precauție la pacienții cu risc de pneumonie de aspirație.
Lactoză: pacienții cu afecțiuni ereditare rare de intoleranță la galactoză, de deficit de lactază lapp sau de malabsorbție la glucoză-galactoză, nu trebuie să utilizeze acest medicament.
28 Hipersensibilitate: similar altor medicamente, în timpul utilizării de aripriprazol pot să apară reacții de hipersensibilitate, reacții caracterizate prin simptome alergice (vezi pct 4. 8).
4. 5 Interacțiuni cu alte medicamente și alte forme de interacțiune
Deoarece aripiprazolul este un antagonist al receptorilor α1-adrenergici, poate să potențeze efectul anumitor antihipertensive.
4. 8).
Trebuie avut grijă atunci când aripiprazolul este administrat împreună cu alte medicamente cunoscute a determina prelungirea intervalului QT sau a afecta echilibrul electrolitic.
Potențialul altor medicamente de a afecta ABILIFY:
Un inhibitor al secreției gastrice acide, famotidina, antagonist al receptorilor H2, reduce viteza absorbției aripiprazolului, dar acest efect nu este considerat relevant clinic.
Aripiprazolul este metabolizat prin multiple căi metabolice care implică enzimele CYP2D6 și CYP3A4, dar nu și enzimele CYP1A.
De aceea, pentru fumători nu este necesară ajustarea dozei.
Într-un studiu clinic la subiecți sănătoși, un inhibitor potent al CYP2D6 (chinidina) a crescut ASC de aripiprazol cu 107%, în timp ce Cmax a rămas neschimbată.
Valorile ASC și Cmax de dehidro -aripiprazol, metabolitul activ, au scăzut cu 32%, respectiv cu 47%.
În cazul administrării concomitente de ABILIFY cu chinidină, doza deABILIFY trebuie redusă la aproximativ o jumătate din doza prescrisă.
Deoarece se așteaptă ca alți inhibitori potenți ai CYP2D6, cum sunt: fluoxetina și paroxetina, să aibă efecte similare, trebuie aplicate reduceri similare ale dozei.
Într-un studiu clinic la subiecți sănătoși, un inhibitor potent al CYP3A4 (ketoconazolul) a crescut ASC și Cmax de aripiprazol cu 63%, respectiv cu 37%.
Valorile ASC și Cmax de dehidro-aripiprazol au crescut cu 77%, respectiv cu 43%.
La pacienții care metabolizează lent prin CYP2D6, utilizarea concomitentă a inhibitorilor potenți ai CYP3A4 poate determina concentrații plasmatice mai mari de aripiprazol, comparativ cu cele ale pacienților care metabolizează rapid prin CYP2D6.
În cazul în care se are în vedere administrarea concomitentă a ketoconazolului sau a altor inhibitori potenți ai CYP3A4 cu ABILIFY, beneficiile potențiale trebuie să depășească eventualele riscuri pentru pacient.
Dacă se administrează ketoconozol concomitent cu ABILIFY, doza de ABILIFY trebuie redusă la aproximativ jumătate din doza prescrisă.
Deoarece se așteaptă ca alți inhibitori potenți ai CYP3A4, cum sunt: itraconazolul și inhibitorii proteazei HIV, să prezinte efecte similare, trebuie aplicate reduceri similare ale dozei.
După întreruperea administrării unui inhibitor al CYP2D6 sau 3A4, dozele de ABILIFY trebuie crescute la valorile anterioare inițierii terapiei concomitente.
Atunci când ABILIFY este utilizat împreună cu inhibitori slabi de CYP3A4 (de exemplu: diltiazem sau escitalopram) sau de CYP2D6, ar putea apărea o creștere moderată a concentrației aripiprazolului.

După administrarea concomitentă a carbamazepinei, un inductor potent al CYP3A4, mediile geometrice ale valorilor Cmax și ASC de aripiprazol au fost cu 68%, respectiv cu 73% mai mici, comparativ cu valorile obținute în cazul administrării aripiprazolului (30 mg) în monoterapie.
În mod similar, după administrarea concomitentă a carbamazepinei, pentru dehidro-aripiprazol, mediile geometrice ale valorilor Cmax și ASC au fost cu 69%, respectiv cu 71% mai mici, decât cele obținute după monoterapie cu aripiprazol.
29 În cazul administrării concomitente de ABILIFY cu carbamazepină, doza de ABILIFY trebuie dublată.
Deoarece se așteaptă ca alți inductori potenți ai CYP3A4 (cum sunt: rifampicina, rifabutina, fenitoina, fenobarbitalul, primidona, efavirenzul, nevirapina și sunătoarea) să prezinte efecte similare, trebuie aplicate creșteri similare ale dozei.
După întreruperea administrării inductorilor potenți ai CYP3A4, doza de ABILIFY trebuie redusă la doza recomandată.
Atunci când fie litiul, fie valproatul au fost administrate concomitent cu aripiprazol, nu s-a observat nici o modificare semnificativă clinic a concentrațiilor de aripiprazol.
Potențialul ABILIFY de a afecta alte medicamente:
În studiile clinice, doze de aripiprazol de 10-30 mg pe zi nu au prezentat un efect semnificativ asupra metabolizării substraturilor CYP2D6 (raport dextrometorfan/ 3-metoximorfinan), 2C9 (warfarină), 2C19 (omeprazol) și 3A4 (dextrometorfan).
În plus, in vitro, aripiprazol și dehidro-aripiprazol nu au dovedit potențial de alterare a metabolizării mediate pe calea CYP1A2.
De aceea, este puțin probabil ca aripiprazolul să determine interacțiuni medicamentoase importante din punct de vedere clinic, mediate de către aceste enzime.
Atunci când aripiprazolul s-a administrat concomitent, fie cu valproat, fie cu litiu, nu s-a observat nicio modificare importantă clinic a concentrațiilor de valproat sau de litiu.
4. 6 Sarcina și alăptarea
Nu există studii controlate, adecvate cu aripiprazol la femeile gravide.
5. 3).
Pacientele trebuie sfătuite să-și informeze medicul dacă devin gravide sau intenționează să devină gravide în timpul tratamentului cu aripiprazol.
Din cauza informațiilor insuficiente privind siguranța la om și a problemelor ridicate de studiile privind reproducerea la animale, acest medicament nu trebuie utilizat în timpul sarcinii, cu excepția cazurilor în care beneficiile așteptate justifică clar riscul potențial pentru făt.
Aripiprazolul s-a excretat în laptele femelelor de șobolan tratate.
La om, nu se cunoaște dacă aripiprazolul se excretă în laptele matern.
Pacientele trebuie sfătuite să nu alăpteze dacă utilizează aripiprazol.
4. 7 Efecte asupra capacității de a conduce vehicule și de a folosi utilaje
Nu s-au efectuat studii privind efectele asupra capacității de a conduce vehicule sau de a folosi utilaje.
Cu toate acestea, ca și în cazul altor antipsihotice, pacienții trebuie avertizați să nu folosească utilaje periculoase, inclusiv vehicule motorizate, până nu sunt siguri că aripiprazolul nu-i afectează defavorabil.
4. 8 Reacții adverse
30 Următoarele reacții adverse apar mai frecvent (≥ 1/ 100) decât cu placebo sau au fost identificate ca reacții adverse posibil relevante medical (*):

Frecvențele prezentate mai jos sunt definite utilizând următoarea convenție: frecvente (> 1/ 100, < 1/ 10) și mai puțin frecvente (> 1/ 1000, < 1/ 100).
Tulburări cardiace Mai puțin frecvente: tahicardie * Tulburări ale sistemului nervos Frecvente: tulburări extrapiramidale, acatisie, tremor, amețeli, somnolență/ sedare, cefalee Tulburări oculare Frecvente: vedere încețoșată Tulburări gastro-intestinale Frecvente: dispepsie, vărsături, greață, constipație, hipersecreție salivară Tulburări vasculare Mai puțin frecvente: hipotensiune arterială ortostatică * Tulburări generale și la nivelul locului de administrare Frecvente: oboseală Tulburări psihice:
Frecvente: neliniște, insomnie, anxietate Mai puțin frecvente: depresie *
Simptome extrapiramidale (SEP):
Schizofrenie: într-un studiu controlat pe termen lung, cu durata de 52-săptămâni, pacienții tratați cu aripiprazol au prezentat o incidență globală a SEP mai mică (25, 8%), incluzând parkinsonism, acatisie, distonie și diskinezie, comparativ cu cei tratați cu haloperidol (57, 3%).
Într-un studiu controlat cu placebo pe termen lung, cu durata de 26-săptămâni, incidența SEP a fost de 19% pentru pacienții tratați cu aripiprazol și de 13, 1% pentru pacienții cărora li s-a administrat placebo.
Într-un alt studiu controlat pe termen lung, cu durata de 26-săptămâni, incidența SEP a fost de 14, 8% pentru pacienții tratați cu aripiprazol și de 15, 1% pentru pacienții tratați cu olanzapină.
Episoadele maniacale în afecțiunea bipolară I: într-un studiu controlat de 12 săptămâni, incidența SEP a fost 23, 5% pentru pacienții tratați cu aripiprazol, și 53, 3% pentru pacienții tratați cu haloperidol.
Într-un alt studiu de 12 săptămâni, incidența SEP a fost de 26, 6% la pacienții tratați cu aripiprazol și 17, 6% la cei tratați cu litiu.
În faza de menținere pe termen lung, a unui studiu placebo -controlat de 26 săptămâni, incidența SEP a fost de 18, 2% pentru pacienții tratați cu aripiprazol și 15, 7% pentru pacienții tratați cu placebo.
În studiile placebo controlate, incidența acatisiei la pacienții cu boală bipolară a fost de 12, 1% cu aripiprazol și 3, 2% cu placebo.

Annex 4

Text extret d'un prospecte de paracetamol del web del Ministeri de Sanitat espanyol per a l'avaluació automàtica de la traducció (es)[37]

Qué necesita saber antes de empezar a tomar Paracetamol Kern Pharma
No tome Paracetamol Kern Pharma
- Si es alérgico al paracetamol o a cualquiera de los demás componentes de este medicamento (incluidos en la sección 6).
Advertencias y precauciones
Consulte a su médico, farmacéutico o enfermero antes de empezar a tomar Paracetamol Kern Pharma.
No tomar más cantidad de medicamento que la recomendada en el apartado 3. Cómo tomar Paracetamol Kern Pharma.
Los alcohólicos crónicos, deberán tener la precaución de no tomar más de 2 g/en 24 horas de paracetamol.
Los pacientes con enfermedades del riñón, del hígado, del corazón o del pulmón y los pacientes con anemia, deberán consultar con el médico antes de tomar este medicamento.
Cuando se está en tratamiento con algún medicamento para tratar la epilepsia debe consultar al médico antes de tomar este medicamento, debido a que cuando se usan al mismo tiempo, se disminuye la eficacia y se potencia la hepatotoxicidad del paracetamol, especialmente en tratamientos con dosis altas de paracetamol.
Los pacientes asmáticos sensibles al ácido acetilsalicílico, deberán consultar con el médico antes de tomar este medicamento.

37 URL: https://www.aemps.gob.es/cima/dochtml/p/69429/Prospecto_69429.html

Annex 5

Postedició al romanés del text extret d'un prospecte de paracetamol del web del Ministeri de Sanitat espanyol per a l'avaluació automàtica de la traducció

Ce trebuie să știți înainte de a lua paracetamol Kern Pharma
Nu luați paracetamol Kern Pharma
- Dacă sunteți alergic la paracetamol sau la oricare dintre celelalte componente ale medicamentului (incluse la punctul 6)
Avertizări și precauții
Adresați-vă medicului dumneavoastră, farmacistului sau asistentei medicale înainte să luați paracetamol Kern Pharma.
Nu luați mai mult decât cantitatea de medicament recomandată la secțiunea 3. Cum să luați paracetamol Kern Pharma.
Alcoolicii cronici trebuie să ia măsuri de precauție pentru a evita să ia mai mult de 2 g într-un interval de 24 de ore de paracetamol.
Pacienții cu boli la rinichi, ficat, cardiace sau pulmonare și pacienții cu anemie trebuie să discute cu medicul înainte de a lua acest medicament.
În timpul tratamentului cu orice medicamente pentru tratarea epilepsiei trebuie să vă adresați medicului înainte de a lua acest medicament, pentru că, dacă sunt luate în același timp, se reduce eficacitea și crește hepatotoxicitatea paracetamolului, în special la terapii cu doze mari de paracetamol.
Pacienții astmatici sensibili la acidul acetilsalicilic trebuie să discute cu medicul înainte de a lua acest medicament.

Annex 6

Text extret d'un prospecte d'omeprazol del web Ce se întâmplă doctore *per a l'avaluació automàtica de la traducció (ro)*[38]

Ce este Omeprazol Terapia și pentru ce se utilizează
Omeprazol Terapia conține substanța activă omeprazol. Aparține unei clase de medicamente denumite "inhibitori ai pompei de protoni". Ele acționează prin scăderea cantității de acid produse de stomacul dumneavoastră.
Omeprazol Terapia este folosit pentru tratamentul următoarelor afecțiuni:
La adulți:
„Boala de reflux gastro-esofagian" (BRGE). În această boală, acidul din stomac trece în esofag (tubul care unește gâtul cu stomacul) producând durere, inflamație și senzație de arsuri în capul pieptului (pirozis).
Ulcere în partea de sus a intestinului (ulcer duodenal) sau în stomac (ulcer gastric).
Ulcere infectate cu bacteria numită „Helicobacterpylori". Dacă aveți această afecțiune, medicul dumneavoastră poate să vă prescrie și antibiotice pentru a trata infecția și a permite ulcerului să se vindece.
Ulcere produse de medicamentele numite AINS (Anti-Inflamatoare Non-Steroidiene). Omeprazol Terapia poate fi utilizat și pentru a împiedica formarea ulcerelor atunci când luați AINS.
Prea mult acid în stomac din cauza unei tumori din pancreas (sindromul Zollinger-Ellison).
La copii:
Copii cu vârsta peste 1 an și greutate corporală > 10 kg
- „Boala de reflux gastro-esofagian" (BRGE). În această boală, acidul din stomac trece în esofag (tubul care unește gâtul cu stomacul) producând durere, inflamație și senzație de arsuri în capul pieptului (pirozis).
Printre simptomele acestei tulburări, la copii mai pot apărea și întoarcerea conținutului stomacului în gură (regurgitare), vărsături și creștere prea mică în greutate.
Copii și adolescenți cu vârsta peste 4 ani
Ulcere infectate cu bacteria numită „Helicobacter pylori". Dacă copilul dumneavoastră are această afecțiune, medicul dumneavoastră poate să îi prescrie și antibiotice pentru a trata infecția și a permite ulcerului să se vindece.

38 URL: http://www.csid.ro/medicamente/omeprazol-terapia-20-mg-capsule-gastrorezistente-11474561/

Annex 7

Postedició al castellà del ext extret d'un prospecte d'omeprazol del web Ce se întâmplă doctore *per a l'avaluació automàtica de la traducció*

Qué es Omeprazol Terapia y para qué se utiliza
Omeprazol Terapia contiene el principio activo omeprazol. Pertenece a una clase de medicamentos denominados "inhibidores de la bomba de protones". Actúan reduciendo la cantidad de ácidos producidos por el estómago.
Omeprazol Terapia se utilitza para el tratamiento de las siguientes patologías:
En adultos:
"Enfermedad por reflujo gastroesofágico" (ERGE). En esta enfermedad, el ácido del estómago pasa al esófago (el tubo que conecta la garganta con el estómago) y causa dolor, inflamación y sensación de ardor en el pecho (pirosis).
Úlceras en la parte superior del intestino (úlcera duodenal) o en el estómago (úlcera gástrica).
Úlceras infectadas por una bacteria llamada ·Helicobacter pylori". Si padece esta enfermedad, su médico puede recetarle además antibióticos para tratar la infección y permitir que cicatrice la úlcera.
Úlceras producidas por medicamentos llamados AINEs (antiinflamatorios no esteroideos). Omeprazol Terapia se puede utilizar para prevenir la formación de úlceras si está tomando AINEs.
Exceso de ácido en el estómago debido a un tumor en el páncreas (síndrome Zollinger-Ellison).
En niños:
Niños de más de 1 año de edad y peso > 10 kg
"Enfermedad por reflujo gastroesofágico (ERGE)". En esta enfermedad, el ácido del estómago pasa al esófago (el tubo que conecta la garganta con el estómago) y causa dolor, inflamación y sensación de ardor en el pecho (pirosis).
Entre los síntomas de este trastorno, en los niños también pueden producirse el retorno del contenido gástrico a la boca (regurgitación), vómitos y un aumento de peso insuficiente.
Niños y adolescentes mayores de más de 4 años
Úlceras infectadas por la bacteria llamada "Helicobacter pylori". Si su hijo padece esta afección, su médico puede recetarle antibióticos para tratar la infección y permitir que cicatrice la úlcera.